ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification

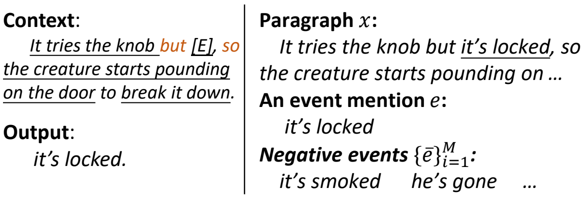
Yucheng Zhou[1*], **Tao Shen**[2], **Xiubo Geng**[2†], **Guodong Long**[1], **Daxin Jiang**[2†]
[1]Australian AI Institute, School of CS, FEIT, University of Technology Sydney
[2]Microsoft
yucheng.zhou.uts@gmail.com, guodong.long@uts.edu.au
{shentao, xigeng, djiang}@microsoft.com

Abstract

Generating new events given context with correlated ones plays a crucial role in many event-centric reasoning tasks. Existing works either limit their scope to specific scenarios or overlook event-level correlations. In this paper, we propose to pre-train a general Correlation-aware context-to-Event Transformer (ClarET) for event-centric reasoning. To achieve this, we propose three novel event-centric objectives, i.e., whole event recovering, contrastive event-correlation encoding and prompt-based event locating, which highlight event-level correlations with effective training. The proposed ClarET is applicable to a wide range of event-centric reasoning scenarios, considering its versatility of (i) event-correlation types (e.g., causal, temporal, contrast), (ii) application formulations (i.e., generation and classification), and (iii) reasoning types (e.g., abductive, counterfactual and ending reasoning). Empirical fine-tuning results, as well as zero- and few-shot learning, on 9 benchmarks (5 generation and 4 classification tasks covering 4 reasoning types with diverse event correlations), verify its effectiveness and generalization ability.

1 Introduction

An 'event', usually a text span composed of a predicate and its arguments (Zhang et al., 2020b), is a fine-grained semantic unit to describe the state of entities/things (e.g., *He looks very worried*) and how they act (e.g., *I grab his arms*). Understanding events and modeling their correlations are fundamental to many reasoning tasks (Bhagavatula et al., 2020; Qin et al., 2019), e.g., abductive reasoning, story ending classification and generation, counterfactual reasoning, script reasoning. For instance, in the left example of Figure 1, to generate the missing event *[E]* in the given context, it is essential to understand that there are four events ('*it tries the knob*', *[E]*, '*the creature starts pounding on the door*', and '*(the creature) to break it down*'), and then predict *[E]* based on the other three events and its correlations to them (i.e., the contrast relation indicated by '*but*' and the causal relation by '*so*').



Context:
It tries the knob but [E], so the creature starts pounding on the door to break it down.

Output:
it's locked.

Paragraph x:
It tries the knob but it's locked, so the creature starts pounding on ...

An event mention e:
it's locked

Negative events $\{\bar{e}\}_{i=1}^{M}$:
it's smoked he's gone ...

Figure 1: *Left*: an example of abductive reasoning which aims to generate the missing event [E] given correlated events (underlined) and connectives (w/ orange) in the context. *Right*: a toy example $(x, e, \{\bar{e}\}_{i=1}^{M})$ of event-rich data for better reading. See Appendix A for real examples to pre-train our model.

Event-aware reasoning has gained much attention and achieved promising success in recent years (Lv et al., 2020; Ding et al., 2019). However, many algorithms are designed to solve only some specific tasks. For example, Qin et al. (2020) propose to improve unsupervised decoding for counterfactual and abductive reasoning; Huang et al. (2021) and Guan et al. (2019) advance story ending generation via incremental encoding and multi-level graph convolutional networks. Although these works show effectiveness in corresponding applications, they are limited to specific scenarios, and cannot generalize well to a broad scope of reasoning.

Meanwhile, some pioneering works follow a recently arising paradigm to conduct event-based pre-training for those downstream reasoning tasks (Yu et al., 2020; Han et al., 2020a; Lin et al., 2020; Zhou et al., 2021b). However, these solutions have their own limitations: COMeT (Hwang et al., 2021) learns event correlations from a human-curated knowledge graph and thus limits its scalability. Han et al. (2020a) and Lin et al. (2020) only model temporal relations and cannot be expanded to other relations (e.g., causal, contrast). EventBERT (Zhou et al., 2021b) is proposed for event-based classifi-

[*] Work is done during internship at Microsoft.
[†] Corresponding author.

cations and is thus inapplicable to generation tasks.

In this work, we propose a general pre-training framework for event-centric reasoning by learning a **C**orre**la**tion-awa**r**e context-to-**E**vent **T**ransformer (ClarET) from an event-rich text corpus. We propose three novel self-supervised objectives, dubbed as whole event recovering (WER), contrastive event-correlation encoding and prompt-based event locating, respectively. The first one aims to capture event correlation by recovering a whole event from its masked context. The second one enhances the representation of the masked event in WER by contrasting it with the gold event against the negative ones. The last one is a simplified WER task by providing hints in its prompt and thus facilitates effective learning for WER.

ClarET explicitly models event correlations and contributes to various scenarios. From one aspect, it covers a variety of correlation types (e.g., causal, temporal, contrast) attributed to correlation type-agnostic objectives. From another aspect, it is applicable to both generation and classification task formulations by its unified structure. Lastly, it highlights event-level correlations and thus is more effective for diverse event-centric tasks, e.g., abductive, counterfactual and ending reasoning.

To evaluate ClarET, we compare it with strong baselines on 9 diverse benchmarks. While ClarET is continually pre-trained from BART (Lewis et al., 2020) with very limited extra resources, i.e., training on a small subset of BART-used corpus (i.e., 200M out of 2.2T tokens) within 90 GPU hours (only 0.13% of 70,000h BART pre-training), it achieves state-of-the-art (SoTA) performance on all 5 generation benchmarks. It also outperforms all unified models on 4 classification benchmarks and achieves competitive, or even better, accuracy to strong discriminative baselines. We further exhibit that the ClarET provides a good initialization for downstream tasks by zero- and few-shot learning.

2 Related Work

Unified Pre-trained Model. A recent trend is to pre-train unified (a.k.a. universal or general) models to boost downstream generation and classification tasks, rather than masked language modeling (MLM) only. GPT (Radford et al., 2019) is based on auto-regressive language modeling but incompetent in classifications due to unidirectional contextualizing. To remedy this, BART (Lewis et al., 2020) trains seq2seq models as a text denoising autoencoder with mask-infilling, etc; UniLM (Dong et al., 2019) designs advanced self-attention masks in Transformer, leading to a partially auto-regressive MLM; GLM (Du et al., 2021) proposes an auto-regressive blank-filling objective based on Transformer, achieved by bi-/uni-directional attention and 2D positional encoding. T5 (Raffel et al., 2020) pre-trains a text-to-text Transformer to recover the masked part of input by decoding. All these general-purpose pre-trained models focus on relatively short-span masking in random, whereas we focus on masking a whole semantic unit (i.e., event) and propose novel training objectives to circumvent problems in long-span event decoding. Besides, they are also vulnerable to pretrain-finetune inconsistency, leading to inferior event-centric performance.

Task-specific Models for Event Reasoning. Many recent works present task-specific neural models for various event-centric reasoning types, including (1) abductive reasoning (Ji et al., 2020; Dong et al., 2021; Zhu et al., 2020), (2) counterfactual reasoning (Qin et al., 2019, 2020), (3) ending reasoning (Guan et al., 2019; Wang and Wan, 2019; Yao et al., 2019; Huang et al., 2021; Guan et al., 2020; Wang et al., 2017; Li et al., 2018; Ding et al., 2019; Zhou et al., 2021c; Chaturvedi et al., 2017; Srinivasan et al., 2018), (4) incoherence reasoning (Mori et al., 2020). However, these methods are designed for the specific reasoning scenarios based on task-specific models so hardly generalize to other scenarios. In contrast, we aim to pre-train a general event-centric model for generalizing to various scenarios.

Event-centric Pre-training. With similar scopes, many works focus on event-centric pre-training to promote event-related tasks as 'event' is a self-contained semantic unit and also an entry of commonsense reasoning. One paradigm is to pre-train on corpora without human-labeling. Some methods focus on more specific aspects of events and their correlations. DEER (Han et al., 2020b) performs temporal and event masking predictions for temporal relations. Lin et al. (2021) propose to recover a temporally-disordered or event-missing sequence for temporal and causal relations. Wang et al. (2021) use AMR structure to design contrastive objectives for the event detection task. However, they are not general enough to various event reasoning tasks. In contrast, CoCoLM (Yu et al., 2020) learns an event-level MLM to generalize more. EventBERT (Zhou et al., 2021b) states the ineffectiveness of event-level MLM and exploits hard negatives via contrasting, contributing much to downstream multi-choice tasks. However, these methods are only competent in discriminative tasks. The other paradigm is based on supervised pre-training on similar tasks and then performs knowledge transfer, e.g., COMeT (Hwang et al., 2021), UnifiedQA (Khashabi et al., 2020) and UNICORN (Lourie et al., 2021), but they require human-curated data.

Event-rich Corpus. Although raw corpora are viewed as off-the-shelf pre-training resources, a key question is how to mine event-rich examples. Here, 'event-rich' denotes that each example contains various events and entails adequate contexts to support event reasoning via either explicit or implicit event-correlation. This is crucial to learning event-correlations and reducing unnecessary overheads. Except for human-curated resources (e.g., ATOMIC (Sap et al., 2019) and ConceptNet (Speer et al., 2017)), event-rich corpora are also mined via automatic schemes. ASER (Zhang et al., 2020b) builds an event-based graph, where each node is an event extracted from a text and the relation of an event pair is predicted by a PDTB model. In contrast, EventBERT (Zhou et al., 2021b) operates on pure text so filters out correlation-scarce contexts and extracts verb-rooted events. Besides, it offers event sampling methods for hard negatives. We adopt this data processing method as both pure-text examples and hard negatives are prerequisites of generic and robust pre-training.

3 Methodology

3.1 Prerequisite: Event-rich Corpus

In this work, we directly adopt event-rich data mining and negative sampling methods from Zhou et al. (2021b) but focus our contributions on enlarging application scope of event-centric tasks and overcoming challenges raised in the new scope.

Event-rich Data Mining. To mine event-rich data from raw corpus, we employ a story corpus, BOOKCORPUS (Zhu et al., 2015), and take a two-step procedural (i.e., '*filter*' and '*extraction*'). It *filters* out correlation-scarce paragraphs according to existence of connectives (i.e., discourse relation keywords, e.g., *however*, *while*). Then, it highlights the event spans in the filtered paragraphs by *extracting* verb-rooted sub-trees in dependency trees of the paragraphs. With a filtered paragraph x, we build each example as (x, e) where e is an event mention in x. We obtain 200M tokens (out of 1B in BOOKCORPUS) in 3.9M filtered paragraphs. ***For clear notations***, we denote a text piece as a lower case letter (e.g., e). It is tokenized into a sequence as a bold (e.g., $\boldsymbol{e} = [e_1, e_2, \dots]$), where a letter w/ subscript t is the t-th token in the sequence.

Negative Event Sampling. Following Zhou et al. (2021b), we build a pool of events from the whole corpus and then retrieve negative events by three heuristic schemes. Given an event e in (x, e), we sample its negative event, $\bar{e}$, in light of lexicon-based (20% time), PoS-based (60% time) or in-domain (20% time) retrieval. Consequently, given an event e, we sample M negative events, i.e., $\{\bar{e}\}_{i=1}^{M}$. Figure 1 (right) shows an integrated instance $(x, e, \{\bar{e}\}_{i=1}^{M})$ of the event-rich corpus[1].

3.2 Pre-training Objectives

We first present *whole event recovering* as a backbone pre-training objective in §3.2.1. After identifying incompetence of the simple backbone, we propose two other objectives in §3.2.2 and §3.2.3. An overview of the objectives is shown in Figure 2.

3.2.1 Whole Event Recovering

For the objective of whole event recovering (WER), it is straightforward to leverage an encoder-decoder structure, where a masked context is passed into the encoder to generate the missing part by decoding. Specifically, given an event e in a paragraph x, we mask out e from x at the encoder side and then generate e at the decoder side, i.e.,

$$p(e|x_{/\{e\}};\theta) = \prod\nolimits_t p(e_t|e_{<t}, x_{/\{e\}};\theta), \quad (1)$$

where θ denotes parameters and $x_{/\{e\}}$ denotes replacing e in x with *one* special token `[M]`. We estimate Eq. (1) by the Transformer sequence-to-sequence (seq2seq) structure (Vaswani et al., 2017). First, we apply the Transformer encoder to $x_{/\{m\}}$ for contextual embeddings for all tokens in $x_{/\{m\}}$:

$$\boldsymbol{H}^{(enc)}=\text{Trans-Enc}(x_{/\{e\}};\theta^{(enc)}) \in \mathbb{R}^{d\times n}, \quad (2)$$

where n is the number of tokens in $x_{/\{e\}}$. Then, the Transformer decoder is employed to predict all

[1] Codes are uploaded to the submission system, which will be released with the mined corpus and pre-trained models.

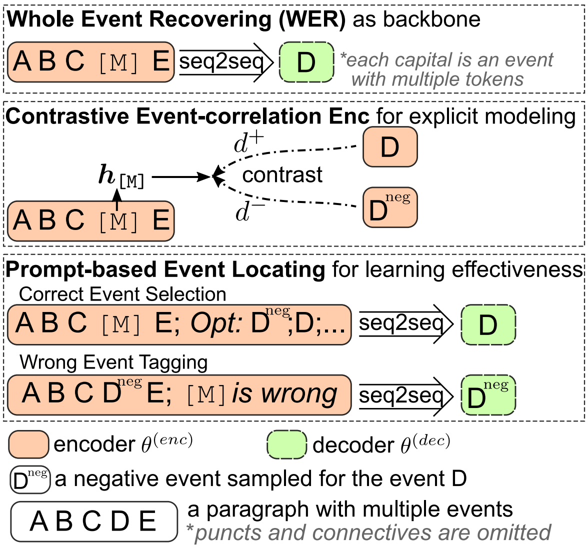

Whole Event Recovering (WER) as backbone
A B C [M] E
seq2seq
D
*each capital is an event with multiple tokens
Contrastive Event-correlation Enc for explicit modeling
d^+
D
$h_{[M]}$
contrast
A B C [M] E
d^-
D^{neg}
Prompt-based Event Locating for learning effectiveness
Correct Event Selection
A B C [M] E; Opt: D^{neg};D;...
seq2seq
D
Wrong Event Tagging
A B C D^{neg} E; [M] is wrong
seq2seq
D^{neg}
encoder $\theta^{(enc)}$
decoder $\theta^{(dec)}$
D^{neg} a negative event sampled for the event D
A B C D E
a paragraph with multiple events
*puncts and connectives are omitted

Figure 2: An overview of self-supervised objectives for our **C**orre**la**tion-awa**r**e context-to-**E**vent **T**ransformer (ClarET).

tokens $\boldsymbol{e}$ of the event e in a recurrent manner, i.e.,

$$\tilde{\boldsymbol{y}}_t = \text{Trans-Dec}(e_{<t}, \boldsymbol{H}^{(enc)}; \theta^{(dec)}) \in \mathbb{R}^{|\mathcal{V}|}, \quad (3)$$

where $\mathcal{V}$ denotes token vocabulary and $\tilde{\boldsymbol{y}}_t$ is the predicted categorical distribution over $\mathcal{V}$. Lastly, the training objective is defined as a maximum likelihood estimation. Its loss function is written as

$$L^{(wer)} = -\sum_{(x,e)} \frac{1}{|\boldsymbol{e}|} \sum\nolimits_{t=1}^{|\boldsymbol{e}|} \log \tilde{\boldsymbol{y}}_t[y = e_t], \quad (4)$$

where '$\boldsymbol{y}_t[y = e_t]$' denotes fetching the probability of the t-step gold token $e_t \in \boldsymbol{e}$ from $\tilde{\boldsymbol{y}}_t$.

This objective is similar to span recovering schema (Raffel et al., 2020; Joshi et al., 2020) but differs in that (i) each masked span is an event, i.e., an integrated semantic unit, so much longer (up to 22 tokens and see Figure 4 for length distribution), and (ii) only one event is masked out from the context to facilitate event-correlation modeling between the event and its contexts.

Intuitively, the success of Eq. (1) requires to capture correlations between the masked event and remaining contexts but two major problems arise due to WER with long event-level masking spans:

(1) Implicit Event-correlation: The model recovers an event based solely on token-level concurrence as in a conditional language model (e.g., T5 and BART), regardless of the rich event-level correlations between the events in context $x_{/\{e\}}$ and the masked event e. Such a correlation-implicit model would achieve inferior performance on downstream event-centric correlation reasoning tasks.

(2) Learning Difficulty: As the masked event is an integrated, self-contained, semantic unit, it is difficult for the conditional generation model to recover the whole event due to a lack of local contexts. As a result, the model cannot effectively learn from the long masked spans, which has been empirically proved in autoencoding MLM models.

To alleviate the two problems above, we propose two other novel self-supervised objectives in the following. Briefly, we present contrastive event-correlation encoding to enhance correlations between contexts and events, and prompt-based event locating to reduce generation difficulty.

3.2.2 Contrastive Event-correlation Encoding

For the *implicit event-correlation* problem, an intuitive solution is to explicitly highlight the correlation from the masked context to the missing event at the encoder side. To achieve this, we resort to contrastive learning to enhance the encoder-side representation of the masked event by contrasting it with the embedding of the gold event mention e against those of negative ones $\bar{e}$. Particularly, we first derive the embedding of e and $\bar{e}$ independently via the Transformer encoder in Eq.(2), i.e.,

$$\boldsymbol{c} = \text{Pool}(\text{Trans-Enc}(\texttt{[CLS]} + e; \theta^{(enc)})), \quad (5)$$
$$\bar{\boldsymbol{c}} = \text{Pool}(\text{Trans-Enc}(\texttt{[CLS]} + \bar{e}; \theta^{(enc)})), \quad (6)$$

where `[CLS]` is a special token prefixed to each event mention, and $\text{Pool}(\cdot)$ denotes using the contextual embedding of `[CLS]` to represent the whole event. Then, we enhance $\boldsymbol{h}_{\texttt{[m]}}$, the contextual representation of `[M]` in $x_{/\{e\}}$ from $\boldsymbol{H}^{(enc)}$ in Eq.(2), by contrasting it with $\boldsymbol{c}$ against $\bar{\boldsymbol{c}}$, i.e.,

$$L^{(cee)} = \max(0, \lambda + d(\boldsymbol{h}_{\texttt{[m]}}, \boldsymbol{c}) - d(\boldsymbol{h}_{\texttt{[m]}}, \bar{\boldsymbol{c}})), \quad (7)$$

where $d(\cdot, \cdot)$ denotes a distance metric of two vectors, which is Euclidean distance in this work. As a result, the encoder-side correlation-aware representation $\boldsymbol{h}_{\texttt{[m]}}$ also offers a straightforward pathway to transmit event-level information to decoding so mitigates the *learning difficulty* to some extent.

3.2.3 Prompt-based Event Locating

As for *learning difficulty* problem, we also propose a prompt-based event locating objective to reduce generative difficulty by providing hints in the prompt. The basic idea is to simplify WER objective as an extractive generation task to locate and copy a candidate/hint from the prompt, which aims at improving learning effectiveness. To this end, we present two prompt-based generation schemas in the following.

Correct Event Selection. Inspired by advances of prompt-based multi-choice question answering, we present correct event selection schema to select the gold event e against negative ones $\{\bar{e}\}_{i=1}^{M}$ based on the contexts $x_{/\{e\}}$. Given an event-masked paragraph $x_{/\{e\}}$ suffixed with several candidate events $\{\bar{e}\}_{i=1}^{M}$ containing the gold masked one e, it aims to generate the masked event e back, i.e.,

$$\hat{x}^{(ces)} = x_{/\{e\}} + \underline{\text{Options: (a) } e^1\text{; (b) } e^2\text{; } \cdots},$$

where $[e^1, e^2, \ldots]$ is a random permutation of $[e, \{\bar{e}\}_{i=1}^{M}]$ in case of position bias. We use a random permutation as all candidates are assigned with distinct position embeddings during contextualizing, and a fixed permutation of gold events will result in a learning shortcut (position bias) to degrade the model. Thus, similar to Eq.(1), we can define its formula as $p(e|\hat{x}^{(ces)}; \theta)$.

Wrong Event Tagging. The other schema is wrong event tagging to find the wrong event in a corrupted paragraph, similar to incoherence reasoning. Thus, we re-write the encoder input as

$$\hat{x}^{(wet)} = x_{/\{e\}\&\cup\{\bar{e}\}} + \underline{\text{Event: } \texttt{[M]} \text{ is wrong}},$$

where $x_{/\{e\}\&\cup\{\bar{e}\}}$ denotes replacing the gold event e in x with a negative $\bar{e} \in \{\bar{e}\}_{i=1}^{M}$. Thus, we can define the formula of this objective as $p(\bar{e}|\hat{x}^{(wet)}; \theta)$.

Based on the two formulas above, we define the prompt-based event locating objective as

$$L^{(pel)} = \sum_{(x,e)} -\frac{1}{|\mathbf{e}|} \sum_{t} \log p(e_t|e_{<t}, \hat{x}^{(ces)}; \theta) - \frac{1}{|\bar{\mathbf{e}}|} \sum_{t} \log p(\bar{e}_t|\bar{e}_{<t}, \hat{x}^{(wet)}; \theta), \quad (8)$$

where $\theta = \{\theta^{(enc)}, \theta^{(dec)}\}$, $\bar{e}$ is sampled in $\{\bar{e}\}_{i=1}^{M}$.

3.3 Model Pre-training and Fine-tuning

Self-supervised Pre-training. The final loss to pre-train our ClarET is a linear combination of the three losses above from Eq.(4, 7, 8), i.e.,

$$L = L^{(wer)} + L^{(cee)} + L^{(pel)}. \quad (9)$$

We set the margin λ in Eq.(7) to 0.5 w/o tuning.

Supervised Downstream Fine-tuning. For generation tasks, we simply leverage the formula in Eq.(1) to establish fine-tuning objectives. For discriminative (e.g., multi-choice) tasks, we can either formulate all tasks into generation as in GPT/T5 or fine-tune with classifying heads as in BART. With pilot experiments, we found the latter one can achieve better performance and adopted it.

3.4 Comparing to Similar Works

While we adopt the same data processing in EventBERT (Zhou et al., 2021b) and share a similar motivation to learn an event-centric pre-trained model, we expand the scope from '*discriminative-only*' in EventBERT into '*unified*' by our context-to-event Transformer for a broad spectrum of scenarios. Such an expansion is non-trivial since new challenges arise in the unified formulation. Compared to the inefficient '*event-backfilling and contextualizing*' paradigm in EventBERT, our model can explicitly and effectively learn event-level correlations between contexts and events by our novel contrastive and prompt-based objectives. Moreover, COMeT (Bosselut et al., 2019; Hwang et al., 2021) is also a conditional generation model but focuses on triple-level commonsense reasoning – given (*head event*, *relation*) to generate *tail events*, whose motivation, however, is orthogonal to ours. Therefore, we focus on a different motivation or scope, not to mention evaluation formulations.

4 Experiments

This section begins with descriptions of downstream datasets and experimental setups.

Downstream Datasets. We conduct extensive evaluations on 9 datasets for 9 downstream tasks, i.e., 5 generation and 4 classification tasks. Generation tasks include abductive commonsense reasoning on $\mathcal{ART}$ (αNLG) (Bhagavatula et al., 2020), counterfactual story generation on TIMETRAVEL (Qin et al., 2019), story ending generation (Guan et al., 2019), commonsense story generation (Guan et al., 2020), and event process completion on APSI (Zhang et al., 2020a). Classification tasks include script reasoning on MCNC (Li et al., 2018), abductive commonsense reasoning on $\mathcal{ART}$ (αNLI) (Bhagavatula et al., 2020), narrative incoherence detection on ROCStories (Mori et al., 2020), and story cloze test (Mostafazadeh et al., 2016). Please refer to Appendix C for their details.

		Abductive C.S. Reasoning			Counterfactual Story			Story Ending Generation		C.S. Story Generation		Event Process Completion	
	Size	**B-4**	**R-L**	**BERT**	**B-4**	**R-L**	**BERT**	**B-1**	**B-2**	**B-1**	**B-2**	**B-1**	**B-2**
Selected task-specific models with competitive performance													
GRF (Ji et al., 2020)	-	11.62	34.62	-	-	-	-	-	-	-	-	-	-
IE+MSA (Guan et al., 2019)	-	-	-	-	-	-	-	24.40	7.80	-	-	-	-
Plan&Write (Yao et al., 2019)	-	-	-	-	-	-	-	24.40	8.40	30.80	12.60	-	-
Fine-tuning with pre-trained unified (generative) model													
GPT2-S (Radford et al., 2019)	124M	2.23	22.83	48.74	69.27	65.72	60.53	39.23	13.08	32.20	14.10	35.25	11.75
GPT2-M (Radford et al., 2019)	335M	-	-	-	75.71	72.72	62.39	-	-	-	-	45.43	14.81
BART (Lewis et al., 2020)	400M	16.47	38.73	56.36	82.91	76.44	79.50	54.22	18.07	54.22	18.07	56.25	18.75
GLM (Du et al., 2021)	335M	7.79	25.54	54.85	75.81	70.03	68.23	57.04	18.45	57.04	18.45	57.34	19.11
ClarET (ours)	400M	**17.67**	**41.04**	**57.31**	**87.18**	**80.74**	**81.48**	**57.47**	**19.16**	**57.47**	**19.16**	**58.88**	**19.74**

Table 1: Fine-tuning results on five generation benchmark datasets. Previous state-of-the-art (SoTA) results are underlined, 'Size' denotes the number of model parameters, and 'C.S.' is an abbreviation of CommonSense. Please refer to Appendix D.1 for the reported results of more task-specific models on each dataset.

		Abductive C.S. Reasoning	Script Reasoning	Narrative Incoherence Detection	Story Cloze Test
	Size	**ACC (%)**	**ACC (%)**	**ACC (%)**	**ACC (%)**
Selected task-specific models with competitive performance					
Hidden Coherence Model (Chaturvedi et al., 2017)	-	-	-	-	77.60
GRU Context (Mori et al., 2020)	-	-	-	52.20	-
RoBERTa + Kown. Model (Zhou et al., 2021c)	469M	-	63.62	-	-
Fine-tuning with pre-trained discriminative model					
RoBERTa (Liu et al., 2019)	345M	82.35	61.53	73.94	87.10
EventBERT (Zhou et al., 2021b)	345M	85.51	63.50	75.03	91.33
Fine-tuning with pre-trained unified model					
CALM (Zhou et al., 2021a)	770M	77.12	-	-	-
UNICORN (Lourie et al., 2021)	770M	79.50	-	-	-
BART (Lewis et al., 2020)	400M	80.74	61.34	72.48	87.01
ClarET (ours)	400M	**82.77**	**64.61**	**74.88**	**91.18**

Table 2: Fine-tuning results on four classification benchmark datasets. We split pre-trained models into discriminative and unified groups since discriminative models usually outperforms unified ones in classification and our ClarET falls into the latter. Previous SoTA discriminative and unified results are waved and underlined, respectively. See Appendix D.2 for full results.

Pre-training Setups. Instead of learning from scratch, we perform continual pre-training from BART-large (Lewis et al., 2020) due to limited computation resources. The batch size and number of training steps are 1152 and 160k. The model is trained by Adam (Kingma and Ba, 2015) w/ learning rate of 1e-5 and warmup proportion of 0.03. The gradient clip, dropout rate and weight decay are 1.0, 0.1 and 0.01. Notably, (i) BOOKCORPUS has already been used by BART pre-training and our data processing is based on heuristics without human-curated resources; (ii) Our continual pre-training only needs 90 GPU hours on 200M tokens, i.e., 0.13% of BART that consumes 70K hours on 2.2T tokens (see Appendix B.1). Hence, ClarET with zero newly introduced corpus and relatively negligible computing overhead makes great lifts and preserves fair comparisons with baselines.

Fine-tuning Setups. For finetuning, we train the model with an Adam w/ learning rate of 1e-5 and warmup proportion of 0.06. The dropout rate, batch size and weight decay are 0.1, 32 and 0.01. For generative downstream tasks, we take BLEU-N (B-N) (Papineni et al., 2002), ROUGE-L (R-L) (Lin, 2004) and BERTScore (BERT) (Zhang et al., 2020c) as the evaluation metrics, while the accuracy (ACC) is taken for classification tasks. Each fine-tuning runs with seeds 2, 10 and 1234, and we evaluate the best dev model on the test set.

Method	B-4	R-L	BERT
GPT (Qin et al., 2019)	1.25	18.26	59.50
GPT2-S (Qin et al., 2019)	1.28	20.27	59.62
GPT2-M (Qin et al., 2019)	1.51	19.41	60.17
Zero-Shot-Ranked (Qin et al., 2020)	2.26	25.81	60.07
BART-large (Lewis et al., 2020)	7.08	30.60	61.58
DeLorean (Qin et al., 2020)	21.35	40.73	63.36
ClarET (ours)	**23.75**	**43.03**	**63.93**

Table 3: Zero-shot results on generative Counterfactual Story.

4.1 Main Evaluation

Fine-tuning for Generation. As shown in Table 1, our proposed ClarET achieves SoTA performance across all generation tasks. For instance, ClarET increases the ROUGE-L score by 2.3 absolute value for abductive reasoning. The superior performance of ClarET on the benchmarks demonstrates that it can model event-level correlation more effectively via few steps of continual pre-training and provide a general solution for a variety of event-centric correlation reasoning tasks.

Fine-tuning for Classification. Table 2 lists results on 4 classification tasks. We find ClarET performs better than all task-specific models and unified pre-trained models with 2%-4% improvement. It achieves competitive accuracy to strong discriminative models, e.g., the gap between ClarET and EventBERT is ∼0.15 for narrative incoherence detection and story cloze test. However, EventBERT is a RoBERTa-based competitor using the identical pre-training corpus. Its pre-training follows "event-backfilling and contextualizing" (similar to multi-choice QA), which has a small gap to downstream classification tasks for strong performance but brings two drawbacks. Firstly, its pre-training is slow due to repeat contextualizing over paragraphs, leading to 5.6× longer GPU hours than ours. In addition, its discriminative paradigm limits it specifically to classifications, regardless of wide generation tasks. The results show ClarET is on par with the discriminative-only EventBERT on classifications. This is non-trivial given the large formulation gap between our generative pre-training objectives and downstream multi-choice-style classification tasks, and attributed to our effective event-correlation learning. In summary, these results show ClarET serves as a unified pre-trained model for event-centric generation and classification tasks.

Method	ACC (%)
Random	20.00
RoBERTa-large (Zhou et al., 2021b)	20.09
DeBERTa-xlarge (Zhou et al., 2021b)	20.31
BART-large (Lewis et al., 2020)	21.72
EventBERT (Zhou et al., 2021b)	30.79
ClarET (ours)	**32.15**

Table 4: Zero-shot results on discriminative Script Reasoning. Note that MLM-style models are evaluated by autoregression-like operation (Zhou et al., 2021b).

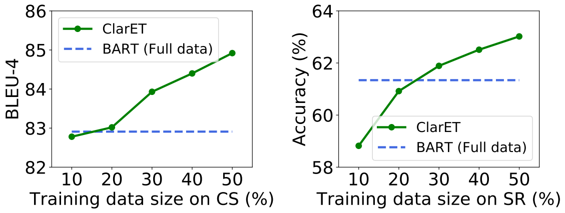


Figure 3: Few-shot learning results compared with the basic model, BART-large, on generation (Counterfactual Story, CS) and classification (Script Reasoning, SR).

4.2 Quantitative Analysis

Zero-shot Learning. It is essential to verify if the targeted information was learned and retained by a pre-trained model. Compared to MLM, our generative recovering model is inherently applicable to event-centric multi-choice and generative formulations. *For generation tasks*, we apply Eq.(1) to generate answers. As shown in Table 3, ClarET achieves the best performance and outperforms DeLorean (which adapts auto-regression for counterfactual reasoning). *For classification tasks*, we apply Eq.(1) to each option for its perplexity and select the option with minimum. As shown in Table 4, ClarET surpasses previous models and beats the discriminative-only event-centric model, EventBERT. Besides, the general-purpose pre-trained models perform nearly random guesses due to their incompetence in long-span event discrimination.

Few-shot Learning. Since our model reduces pretrain-finetune inconsistency for event-centric tasks and provides a good initialization for downstream fine-tuning, it is also interesting to see few-shot performance by scaling down training data. As shown in Figure 3, ClarET achieves similar performance to strong baselines with only 10%-30% of training data for fine-tuning.

Ablation study. To measure the contribution of each objective to the final fine-tuning results, we conduct an ablation study on both generation

Method	Gen-CS		Cls-SR
	B-4	R-L	ACC
ClarET (full, pre-trained by Eq.(9).)	**87.18**	**80.74**	**64.61**
◇ w/o correct event selection (prompt)	86.76	80.03	63.06
◇ w/o wrong event tagging (prompt)	86.33	79.84	63.89
◇ w/o contrastive encoding	85.84	78.69	63.24
◇ only prompt-based event locating	83.32	76.51	62.97
BART-large (basic model)	82.91	76.44	61.34

Table 5: Ablation study of the pre-training objectives in ClarET, which is evaluated by fine-tuning on generation (Counterfactual Story, CS) and classification (Script Reasoning, SR).

Method	ePPL on Dev
ClarET (full model)	8.27
WER-Only Model	8.76

Table 6: Event generation of ClarET and whole event recovering (WER-only) model on a pre-training event-masked dev set (2% held-out masked paragraphs by following Zhou et al. (2021b)). The 'ePPL', i.e., event perplexity, refers to event-level token perplexity averaged over the dataset.

and classification in Table 5. The first two ablations drop the two prompt schemas respectively in prompt-based event locating objective of Eq.(8), which verifies the effectiveness of reducing task difficulty. Then, the third ablation removes contrastive event-correlation encoding and shows a substantial drop, which verifies the significance of explicit event-correlation learning. Next, we keep only the prompt-based event locating objective to make our model a prompt-learning discriminative model (sharing more close methodology with EventBERT), however leading to a dramatic decrease. Lastly, when removing all the objectives, our model degenerates to BART-large.

Comparison with Larger Model. A trend of pre-training models follows the law of 'larger models for better performance' but a crucial research question is 'how to perform competitively with fewer computation resources'. To answer, we show extra fine-tuning results on the five generation datasets in Table 7 to compare our ClarET (400M parameters) with T5-large (770M) and T5-base (220M). It is observed (i) with 3× scale, T5-large notably outperforms T5-base to support the above law and (ii) with almost half model size, our ClarET performs very competitively to T5-large (even better on 3 out of 5 tasks), verifying the significance of our objectives towards event-related knowledge.

		ACR	CS	SEG	CSG	EPC
	Size	R-L	B-4	B-1	B-1	B-1
T5-base	220M	38.40	81.02	52.64	41.28	56.53
T5-large	770M	40.77	**90.62**	57.04	43.82	**59.59**
ClarET	400M	**41.04**	87.18	**57.47**	**48.75**	58.88

Table 7: Fine-tuning generation results to compare with larger pre-trained models. Column names are datasets corresponding to those of Table 1. See Appendix B.2 for full results of T5.

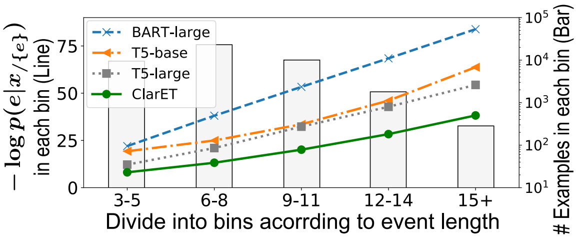


Figure 4: Event generation performance on the event-masked dev set (refer to Table 6) with event-length bins.

Difficulty of Event Generation. To exhibit the *learning difficulty* in pre-training (as stated in §3.2.1) and the effectiveness of our novel learning objectives, we conduct another ablation setting in Table 6. It is observed that ClarET achieves better event-level perplexity (ePPL), verifying the two novel objectives promote event generations and reduce difficulty of decoding.

Long-span Event Generation. To further check if ClarET is more competitive on longer-span event generation, we compare it with BART-large and T5-base/-large by '$-\log$' of Eq.(1). Different from recovering paradigm of others, we follow the denoising paradigm to implement BART and calculate its score by considering the masked part in decoding. Figure 4 shows that *(1) Line Chart:* the gap between ClarET and the others becomes larger with event length increasing as the general-purpose models only consider short-span masking in pre-training, leading to inferior event generation; and *(2) Bar Chart:* as for data distribution, although a majority of data falls into the 6-8 bin, there are still many examples with event length greater than nine.

Natural Language Understanding (NLU). Our basic model, BART-large, is presented for general NLU tasks. To exhibit our minor event-centric continual pre-training would not interfere its NLU ability, we conduct fine-tuning experiments on GLUE benchmark (Wang et al., 2019) as in Figure 5. It is observed that, although slightly surpassed by the discriminative RoBERTa

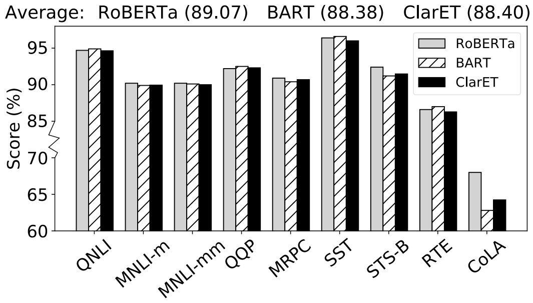


Figure 5: Fine-tuning results on GLUE dev, which verifies ClarET retains BART's natural language understanding ability.

Context: *I went to the store to buy a phone. [E] They were impressed with my phone.*
Reference of the Gold Event *[E]*: *I bought the latest model of the phone I wanted, and showed it to my friends.*
Generation by ClarET:
I bought a new phone and showed it to my friends. (BLEU-4: 34)
Generation by BART:
I bought a new phone. (BLEU-4: 0)

Context: *Cora was starting her job as a kindergarten teacher. [E] At the end of the day, they all told her how much they liked her!*
Reference of the Gold Event *[E]*: *Cora was nervous, but knew the students were nervous too, so she tried to be extra friendly.*
Generation by ClarET:
Cora spent the whole day with her students. (BLEU-4: 0)

Figure 6: Case study & error analysis on abductive reasoning.

model, fine-tuning BART and ClarET achieve very comparable results, which verifies ClarET's retention of NLU capability.

4.3 Case Study and Error Analysis

Case Study. As the first case in Figure 6, we conduct a case study on generative abductive reasoning task, where the fine-tuned ClarET generates an event semantically close to the gold reference, but the BART does not. BART only generates a part of the answer but ignores the event-correlations from '*They were impressed with my phone*', while ClarET completely captures the correlations in the contexts (e.g., '*to buy a phone*' and '*They were impressed*',) and generate a much better result.

Error Analysis and Limitation. The second case in Figure 6 shows that our ClarET is ineffective when the gold event is very complicated. In detail, the model focus only on '*at the end of the day*' to generate '... *spent the whole day* ...' but ignore very subtle contexts, e.g., '*starting her job ... teacher*' and '*they liked her*'. To expand, we found a problem in long-event decoding by pilot experiments. As shown in Figure 7, it is observed that the gap of token-level perplexity between ClarET and WER-only gradually diminishes. This is because the subsequent tokens in an event can be generated on the basis of previous generations on the decoder side, rather than context-aware representations from the encoder side. While a long span is masked, the model can see previous tokens in an event (i.e., $e_{<t}$) in decoding and incline to perform the t-th prediction based on $e_{<t}$ but not $x_{/\{e\}}$, especially with a larger t. As a result, the model would 'cheat' in the generation but learn decoder-side language modeling rather than context-aware representations. In the future, we will exploit this problem. Besides, due to computation resources, we choose the model size with 400M and continual pre-training in 90h, limiting the performance.

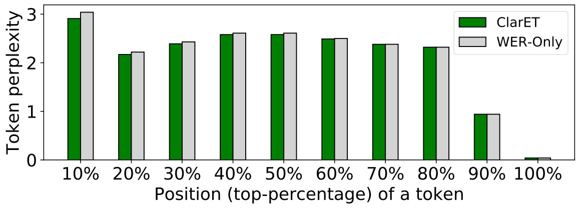


Figure 7: Token-level perplexity w.r.t tokens' percentage positions in events on held-out dev set.

5 Conclusion

We present a novel correlation-aware context-to-event Transformer to self-supervisedly learn event-correlation knowledge from text corpus and benefit various event-centric reasoning scenarios. Besides SoTA fine-tuning results on 5 generation and 4 classification tasks, we conduct zero-/few-shot learning and extensive ablation studies to exhibit our model's effectiveness. Lastly, we find our model is competitive to a twice larger general-purpose model, reduces learning difficulty for event generation, and retains NLU ability from its basic model. Although this work learns context-to-event knowledge, our self-supervised objectives are applicable to other semantically-meaningful text units besides events. For example, text units can be entities and concepts to learn relational and commonsense knowledge, which can benefit more downstream tasks.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Represen-*

tations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1603–1614. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Xiao Ding, Kuo Liao, Ting Liu, Zhongyang Li, and Junwen Duan. 2019. Event representation learning enhanced with external commonsense knowledge. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4893–4902. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Yue Dong, Chandra Bhagavatula, Ximing Lu, Jena D. Hwang, Antoine Bosselut, Jackie Chi Kit Cheung, and Yejin Choi. 2021. On-the-fly attention modulation for neural generation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1261–1274. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. All NLP tasks are generation tasks: A general pretraining framework. *CoRR*, abs/2103.10360.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2650–2660. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2727–2733. AAAI Press.

Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Trans. Assoc. Comput. Linguistics*, 8:93–108.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6473–6480. AAAI Press.

Rujun Han, Xiang Ren, and Nanyun Peng. 2020a. Deer: A data efficient language model for event temporal reasoning. *arXiv preprint arXiv:2012.15283*.

Rujun Han, Xiang Ren, and Nanyun Peng. 2020b. DEER: A data efficient language model for event temporal reasoning. *CoRR*, abs/2012.15283.

Qingbao Huang, Linzhang Mo, Pijian Li, Yi Cai, Qingguang Liu, Jielong Wei, Qing Li, and Ho-fung Leung. 2021. Story ending generation with multi-level graph convolutional networks over dependency trees. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13073–13081. AAAI Press.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 725–736. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1896–1907. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4201–4207. ijcai.org.

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

Shih-Ting Lin, Nathanael Chambers, and Greg Durrett. 2020. Conditional generation of temporally-ordered event sequences. *CoRR*, abs/2012.15786.

Shih-Ting Lin, Nathanael Chambers, and Greg Durrett. 2021. Conditional generation of temporally-ordered event sequences. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7142–7157. Association for Computational Linguistics.

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

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13480–13488. AAAI Press.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Shangwen Lv, Fuqing Zhu, and Songlin Hu. 2020. Integrating external event knowledge for script learning. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 306–315. International Committee on Computational Linguistics.

Yusuke Mori, Hiroaki Yamane, Yusuke Mukuta, and Tatsuya Harada. 2020. Finding and generating a missing part for story completion. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 156–166.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 839–849. The Association for Computational Linguistics.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 43–54. Association for Computational Linguistics.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5042–5052. Association for Computational Linguistics.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 794–805. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 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. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Siddarth Srinivasan, Richa Arora, and Mark Riedl. 2018. A simple and effective approach to the story cloze test. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 92–96. Association for Computational Linguistics.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tianming Wang and Xiaojun Wan. 2019. T-CVAE: transformer-based conditioned variational autoencoder for story completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5233–5239. ijcai.org.

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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 57–67. Association for Computational Linguistics.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021. CLEVE: contrastive pre-training for event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6283–6297. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the*

2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 4306–4315. Association for Computational Linguistics.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph R-CNN for scene graph generation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11205 of *Lecture Notes in Computer Science*, pages 690–706. Springer.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7378–7385. AAAI Press.

Changlong Yu, Hongming Zhang, Yangqiu Song, and Wilfred Ng. 2020. Cocolm: Complex commonsense enhanced language model. *CoRR*, abs/2012.15643.

Hongming Zhang, Muhao Chen, Haoyu Wang, Yangqiu Song, and Dan Roth. 2020a. Analogous process structure induction for sub-event sequence prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1541–1550. Association for Computational Linguistics.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020b. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020c. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, and Xiang Ren. 2021a. Pre-training text-to-text transformers for concept-centric common sense. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yucheng Zhou, Xiubo Geng, Tao Shen, Guodong Long, and Daxin Jiang. 2021b. Eventbert: A pre-trained model for event correlation reasoning. *CoRR*, abs/2110.06533.

Yucheng Zhou, Xiubo Geng, Tao Shen, Jian Pei, Wenqiang Zhang, and Daxin Jiang. 2021c. Modeling event-pair relations in external knowledge graphs for script reasoning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4586–4596. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

Yunchang Zhu, Liang Pang, Yanyan Lan, and Xueqi Cheng. 2020. L2R2: leveraging ranking for abductive reasoning. *CoRR*, abs/2005.11223.

A Examples from Mined Pre-training Corpus

There are some mined pre-training examples shown in Table 8. As in (Zhou et al., 2021b), an example includes a paragraph, events, a selected positive event, connectives of the positive event, and sampled negative events of the positive event.

B More Details

B.1 BART Pre-training Resources

In this section, we analyze BART pre-training resources in terms of text corpora and computation resources.

As for tokens in BART pre-training corpora, BART paper (Lewis et al., 2020) claims using the same corpora as in RoBERTa (Liu et al., 2019) and T5 paper (Raffel et al., 2020) states RoBERTa uses a 2.2T-token text corpus. Thus, we adopt '2.2T' as the number in the main paper.

As for BART pre-training computation overheads, the contributor of BART official code repository said '*We trained for around 11-12 days on 256 gpus.*' at `https://github.com/pytorch/fairseq/issues/1525`, so the BART pre-training takes from 67584 to 73728 GPU hours. Thus, we use '70,000' as the number in the main paper.

	Example 1
Paragraph	It was only months later, when she saw her friend's thin gaunt face, her swollen belly and her quiet desperation, that she had come to her senses. Then she had been filled with a combination of burning rage and deep shame. This had endured over the years undiminished.
Positive Event	she had been filled with a combination of burning rage
Connectives	when; then
Negative Events	he had been loaded with a lot of vampire venom she had been trained in the art of gentler speech I had been blessed with some sort of fire ability he had been transformed into a piece of living statuary he had beened from a block of pale marble I had been circumci sculptsed in the age of infantile apathy ...
	Example 2
Paragraph	Then, when she turned twenty one at the end of last year, she had decided to act on it. A driver's license was something she needed for her business and the identity papers which went with it were needed for a range of other reasons, such as enrolling Catherine for school at the start of this year.
Positive Event	papers which went with it were needed for a range of other reasons
Connectives	then; when; and
Negative Events	bookcases that stood against it had opened like a pair of French doors it had a little bit of magic in it , just for Lizzie it only gave her a place for a couple of days which occasionally crossed a small ridge sometimes of gravel , sometimes of sand that she was going shopping in the city with a couple of other girls publish that proposal in the paper for three weeks ...

Table 8: Some mined pre-training examples.

	Abductive C.S. Reasoning			Counterfactual Story			Story Ending Generation		C.S. Story Generation		Event Process Completion	
	B-4	**R-L**	**BERT**	**B-4**	**R-L**	**BERT**	**B-1**	**B-2**	**B-1**	**B-2**	**B-1**	**B-2**
T5-base	15.65	38.40	55.98	81.02	75.95	79.12	52.64	17.55	41.28	13.76	56.53	18.84
T5-large	17.75	40.77	57.20	90.62	84.03	83.14	57.04	18.45	43.82	14.61	59.59	19.86

Table 9: Full results of T5-base and T5-large on generation tasks.

B.2 Full Results of T5 Model

The full results of T5-base and T5-large on the five generation tasks are shown in Table 9.

B.3 Connectives in Paragraph

As stated by Zhou et al. (2021b), connectives (i.e., discourse relations in the contexts) play important roles to express correlations among events. Therefore, we also find every possible connective r to each (x, e) where r is a connective in x, which immediately links to the verb of e on the parsing tree of x. To leverage the connectives, we also apply the correct event selection in prompt-based event locating objective to r its negatives $\{\bar{r}\}_1^M$, as correct connective selection. Here, $\bar{r}$ is randomly sampled from discourse relations in the PDTB annotation manual (Webber et al., 2019). At 20% times, we use correct connective selection to replace correct event selection in the prompt-based event locating objective.

C Details of Evaluation Datasets

We detail the nine evaluation datasets in the following. The training example in each dataset is shown in Table 10.

- $\mathcal{ART}$ **(αNLG).** Given two observations in natural language, it aims to generate an explicative hypothesis between them. We follow the official data split (Bhagavatula et al., 2020) with 169,654/1,532/3,059 in training/dev/test.
- **TIMETRAVEL.** Given an original story and a counterfactual event, it aims to rewrite the subsequent events to complete a story,

which is compatible with the counterfactual event. We follow the official data split (Qin et al., 2019) with 98,159/5,613/7,484 in training/dev/test.

- **Story Ending Generation.** We evaluate the story ending generation based on ROCStories, which aims to generate a story ending for a given story context. We follow the data split (Guan et al., 2019) with 90,000/4,081/4,081 in training/dev/test.

- **Commonsense Story Generation.** It is based on ROCStories. Given a leading context, it aims to generate a reasonable story. We follow the data split (Guan et al., 2020) with 88,344/4,908/4,909 in training/dev/test.

- **APSI.** We evaluate event process completion on the APSI dataset, where the goal is to generate a subevent for a given event context. We follow the data split (Zhang et al., 2020a) with 13,501/1,316 in training/test.

- **Multi-choice narrative cloze (MCNC).** Given an event chain, it aims to predict the subsequent event from 5 candidates. We follow the data split (Li et al., 2018) with 140,331/10,000/10,000 in training/dev/test.

- $\mathcal{ART}$ **(αNLI).** Given two observations in natural language, it aims to choose the most explicative hypothesis from 2 candidates. We follow the data split (Bhagavatula et al., 2020) with 169,654/1532 samples in training/dev.

- **ROCStories.** We follow (Mori et al., 2020) to use ROCStories for narrative incoherence detection. A random sentence is removed for each five-sentence story, and the goal is to predict the missing position. We follow the data split (Mori et al., 2020) with 78,528/9,816/9,817 in training/dev/test.

- **Story Cloze Test.** Given a 4-sentence context, it aims to select the right ending from two alternative endings. We follow the data split (Mostafazadeh et al., 2016) with 98,161/1,871/1,871 in training/dev/test.

D Detailed Evaluation Results

We detail the full results on nine evaluation datasets as follows.

D.1 Generation Tasks

Generation tasks include abductive commonsense reasoning (αNLG), counterfactual story generation, story ending generation, commonsense story generation, and event process completion. The detailed results of these generation tasks are shown in Table 11, Table 12, Table 13, Table 14, and Table 15, respectively. ClarET achieves state-of-the-art performance on all five generation tasks. In addition, pre-trained language models show their strong generation ability on story generation tasks, i.e., story ending generation and commonsense story generation.

D.2 Classification Tasks

Classification tasks include script reasoning, abductive commonsense reasoning (αNLI), narrative incoherence detection, and story cloze test. The detailed results of these classification tasks are shown in Table 16, Table 17, Table 18, and Table 19, respectively. Compared with unified language models, ClarET achieves state-of-the-art performance. Although strong discriminative models show their great ability on classification tasks, ClarET still achieves competitive performance.

Dataset 1: $\mathcal{ART}$ (αNLG)	
Input	Observation1: The hayride was in October. Observation2: It was the perfect start to the fall season.
Label	Keeping tradition we drank hot cocoa on the ride.
Dataset 2: TIMETRAVEL	
Input	Premise: On my way to work I stopped to get some coffee. Initial: I went through the drive through and placed my order. Original_ending: I paid the cashier and patiently waited for my drink. When she handed me the drink, the lid came off and spilled on me. The coffee hurt and I had to go home and change clothes. Counterfactual: I went inside to place my order.
Label	I paid the cashier and patiently waited at the counter for my drink. When she handed me the drink, the lid came off and spilled on me. The coffee hurt and I had to go home and change clothes.
Dataset 3: Story Ending Generation	
Input	Dan's parents were overweight. Dan was overweight as well. The doctors told his parents it was unhealthy. His parents understood and decided to make a change.
Label	They got themselves and Dan on a diet.
Dataset 4: Commonsense Story Generation	
Input	Carrie had just learned how to ride a bike.
Label	She didn't have a bike of her own. Carrie would sneak rides on her sister's bike. She got nervous on a hill and crashed into a wall. The bike frame bent and Carrie got a deep gash on her leg.
Dataset 5: APSI	
Input	Process name: Treat Pain. Process: Identify cause. learn injury.
Label	Recognize symptom.
Dataset 6: Multi-choice narrative cloze (MCNC)	
Input	Context: compare basketball. buck get basketball. whirl basketball bench. shout out basketball center. Options: A. look basketball. B. weaken basketball. C. throw basketball lot youngster. D. client deny basketball. E. client deny basketball.
Label	C
Dataset 7: $\mathcal{ART}$ (αNLI)	
Input	Observation1: Chad went to get the wheel alignment measured on his car. Observation2: The mechanic provided a working alignment with new body work. Hypothesis1: Chad was waiting for his car to be washed. Hypothesis2: Chad was waiting for his car to be finished.
Label	2
Dataset 8: ROCStories	
Input	Laverne needs to prepare something for her friend's party. She decides to bake a batch of brownies. Laverne tests one of the brownies to make sure it is delicious. The brownies are so delicious Laverne eats two of them.
Label	3
Dataset 9: Story Cloze Test	
Input	Context: Rick grew up in a troubled household. He never found good support in family, and turned to gangs. It wasn't long before Rick got shot in a robbery. The incident caused him to turn a new leaf. Options: A. He is happy now. B. He joined a gang.
Label	A

Table 10: The training examples on different datasets.

Method	**B-4**	**R-L**	**BERT**
GPT2-Fixed (Bhagavatula et al., 2020)	0.00	9.99	36.69
O1-O2-Only (Bhagavatula et al., 2020)	2.23	22.83	48.74
COMeT-T+GPT2 (Bhagavatula et al., 2020)	2.29	22.51	48.46
COMeT-E+GPT2 (Bhagavatula et al., 2020)	3.03	22.93	48.52
Fine-tuned GPT2-L (Dong et al., 2021)	13.52	18.01	-
GRF (Ji et al., 2020)	11.62	34.62	-
BART-large (Lewis et al., 2020)	16.47	38.73	56.36
GLM-large (Du et al., 2021)	7.79	25.54	54.85
ClarET	**17.67**	**41.04**	**57.31**

Table 11: Results on the Abductive Commonsense Reasoning (αNLG).

Method	**B-4**	**R-L**	**BERT**
Human (Qin et al., 2019)	64.93	67.64	61.87
GPT2-S (Radford et al., 2019)	69.27	65.72	60.53
GPT2-M+Rec+CF (Qin et al., 2020)	75.92	70.93	62.49
GPT2-M+Sup (Qin et al., 2020)	75.71	72.72	62.39
BART-large (Lewis et al., 2020)	82.91	76.44	79.50
GLM-large (Du et al., 2021)	75.81	70.03	68.23
ClarET	**87.18**	**80.74**	**81.48**

Table 12: Results on the Counterfactual Story.

Method	**B-1**	**B-2**
Seq2Seq (Luong et al., 2015)	18.50	5.90
Transformer (Vaswani et al., 2017)	17.40	6.00
GCN (Yang et al., 2018)	17.60	6.20
IE+MSA (Guan et al., 2019)	24.40	7.80
T-CVAE (Wang and Wan, 2019)	24.30	7.70
Plan&Write (Yao et al., 2019)	24.40	8.40
MGCN-DP (Huang et al., 2021)	24.60	8.60
GPT2-S (Radford et al., 2019)	39.23	13.08
BART-large (Lewis et al., 2020)	54.22	18.07
ClarET	**57.47**	**19.16**

Table 13: Results on the Story Ending Generation.

Method	**B-1**	**B-2**
ConvS2S (Gehring et al., 2017)	31.20	13.20
Fusion (Fan et al., 2018)	32.20	13.70
Plan&Write (Yao et al., 2019)	30.80	12.60
SKRL (Xu et al., 2018)	26.70	8.80
DSRL (Fan et al., 2019)	29.30	11.70
GPT2-S (Guan et al., 2020)	32.20	14.10
KE-GPT2 (Guan et al., 2020)	32.60	14.30
BART-large (Lewis et al., 2020)	45.24	15.08
ClarET	**48.75**	**16.25**

Table 14: Results on the Commonsense Story Generation.

Method	**B-1**	**B-2**
GPT2-S (Radford et al., 2019)	35.25	11.75
GPT2-M (Radford et al., 2019)	45.43	14.81
BART-large (Lewis et al., 2020)	56.25	18.75
GLM-large (Du et al., 2021)	57.34	19.11
ClarET	**58.88**	**19.74**

Table 15: Results on the Event Process Completion.

Method	**ACC**
Discriminative Model	
Random	20.00
Event-Comp (Granroth-Wilding and Clark, 2016)	49.57
PairLSTM (Wang et al., 2017)	50.83
SGNN (Li et al., 2018)	52.45
SGNN + Int&Senti (Ding et al., 2019)	56.03
RoBERTa-base (Liu et al., 2019)	56.23
RoBERTa-large (Liu et al., 2019)	61.53
RoBERTa + Rep. Fusion (Lv et al., 2020)	58.66
EventBERT (Zhou et al., 2021b)	63.50
RoBERTa + Kown. Model (Zhou et al., 2021c)	63.62
Unified Model	
BART-large (Lewis et al., 2020)	61.34
ClarET	**64.61**

Table 16: Results on the Script Reasoning.

Method	**ACC**
Discriminative Model	
Random	50.00
BERT-base (Devlin et al., 2019)	61.88
ERNIE (Zhang et al., 2019)	63.04
KnowBERT (Peters et al., 2019)	63.18
BERT-large (Devlin et al., 2019)	66.75
RoBERTa-large (Liu et al., 2019)	82.35
EventBERT (Zhou et al., 2021b)	**85.51**
Unified Model	
T5-base (Raffel et al., 2020)	61.10
T5-large (Raffel et al., 2020)	77.80
BART-large (Lewis et al., 2020)	80.74
GLM-large (Du et al., 2021)	65.27
CALM-large (Zhou et al., 2021a)	77.12
UNICORN (Lourie et al., 2021)	79.50
ClarET	82.77

Table 17: Results on the Abductive Commonsense Reasoning (αNLI).

Method	**ACC**
Discriminative Model	
Random	20.00
RoBERTa-large (Liu et al., 2019)	73.94
Max-pool Context (Mori et al., 2020)	35.00
GRU Context (Mori et al., 2020)	52.20
EventBERT (Zhou et al., 2021b)	**75.03**
Unified Model	
BART-large (Lewis et al., 2020)	72.48
ClarET	74.88

Table 18: Results on the Narrative Incoherence Detection.

Method	**ACC**
Discriminative Model	
Random	50.00
Hidden Coherence Model (Chaturvedi et al., 2017)	77.60
val-LS-skip (Srinivasan et al., 2018)	76.50
RoBERTa-large (Liu et al., 2019)	87.10
EventBERT (Zhou et al., 2021b)	**91.33**
Unified Model	
Finetuned Transformer LM (Radford et al., 2018)	86.50
BART-large (Lewis et al., 2020)	87.01
ClarET	91.18

Table 19: Results on the Story Cloze Test.