

# ACCENT: An Automatic Event Commonsense Evaluation Metric for Open-Domain Dialogue Systems

Sarik Ghazarian<sup>1\*</sup> Yijia Shao<sup>2\*†</sup> Rujun Han<sup>3‡</sup> Aram Galstyan<sup>1</sup> Nanyun Peng<sup>4</sup>

<sup>1</sup>University of Southern California / Information Sciences Institute

<sup>2</sup>Peking University

<sup>3</sup>AWS AI Labs

<sup>4</sup>Computer Science Department of University of California, Los Angeles

{sarik, galstyan}@isi.edu, shaoyj@pku.edu.cn, rujunh@amazon.com, violetpeng@cs.ucla.edu

## Abstract

Commonsense reasoning is omnipresent in human communications and thus is an important feature for open-domain dialogue systems. However, evaluating commonsense in dialogue systems is still an open challenge. We take the first step by focusing on *event commonsense* that considers events and their relations, and is crucial in both dialogues and general commonsense reasoning. We propose **ACCENT**, an event commonsense evaluation metric empowered by commonsense knowledge bases (CSKBs). ACCENT first extracts event-relation tuples from a dialogue, and then evaluates the response by scoring the tuples in terms of their compatibility with the CSKB. To evaluate ACCENT, we construct the first public event commonsense evaluation dataset for open-domain dialogues. Our experiments show that ACCENT is an efficient metric for event commonsense evaluation, which achieves higher correlations with human judgments than existing baselines.

## 1 Introduction

Open-domain dialogue systems aim to have natural and engaging conversations with users (Chen et al., 2017). The abundance of dialogue corpus (Dziri et al., 2018) and the development of neural models (Radford et al., 2019; Lewis et al., 2020) enable open-domain dialogue systems to generate grammatically correct and meaningful responses (Zhang et al., 2020d; Bao et al., 2021; Ghazarian et al., 2021). Despite the success, systems still struggle to consistently produce commonsense-compliant responses as humans do. As shown in Figure 1 Example A, the generated response is not compliant with commonsense since “need

\* Equal contribution

† The work was done while the author was conducting a summer internship at UCLA.

‡ The collaboration started when the author was a graduate student at USC.

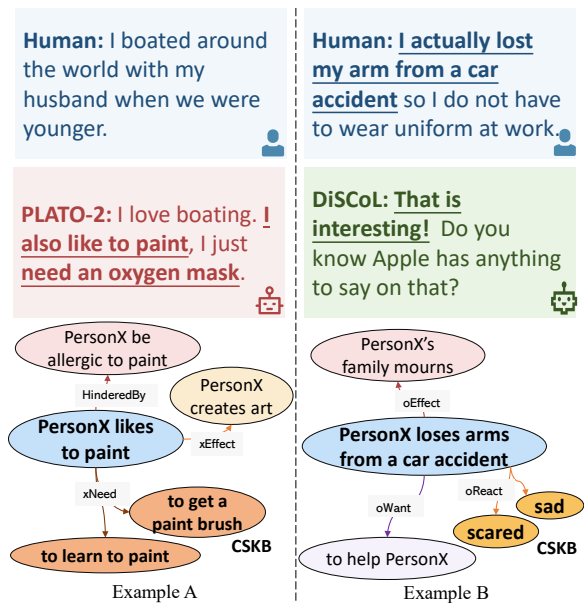


Figure 1: Examples of nonsensical system responses in open-domain dialogues.

an oxygen mask” is not a reasonable prerequisite for “like to paint”. Commonsense issues for dialogue systems can also be manifested when we consider the dialogue history. For instance, in Figure 1 Example B, the system’s response “That is interesting!” after the user talks about their car accident violates commonly accepted social norms (Frischmann, 2021).

In this work, we study automatic dialogue commonsense evaluation by focusing on **event commonsense** (Sap et al., 2020; Rashkin et al., 2018), which concerns commonsense knowledge about events and their relations. Our focus on event commonsense is motivated by the following three observations: First, advanced open-domain dialogue systems have been pre-trained on large corpus and thus suffer less from factoid commonsense issues (Petroni et al., 2019). Second, events and their relations are key components of commonsense reasoning (McCarthy and Hayes, 1981), and our

study shows overall commonsense and event commonsense are highly correlated (see §4). Third, event commonsense aligns well with the interactive nature of open-domain dialogue systems (Huang et al., 2020) to complete certain social goals.

To automatically evaluate event commonsense in open-domain dialogues, we propose **ACCENT**, a reference-free **AutomatiC Event Commonsense EvaluationN meTric** which leverages commonsense knowledge bases (CSKBs) and measures the quality of generated responses without having ground-truth reference responses. For example, comparing the examples in Figure 1 against the CSKB easily reveals commonsense errors in the responses because when “PersonX likes to paint”, what he/she needs may be “to get a paint brush” instead of “to get an oxygen mask”, and when “PersonX loses arms from a car accident”, the other person is expected to feel “sad”.

While these judgments are intuitive to human, two challenges exist in automating the evaluation process. First, there is a considerable gap between free-form conversational data and the compact commonsense knowledge in the CSKB. Second, locating relevant knowledge in the CSKB is non-trivial.

ACCENT addresses these challenges through a pipeline method that uses an intermediate **symbolic representation** for commonsense reasoning. ACCENT first extracts event-relation tuples from the target response and its preceding dialogue history via a prompt-based generative model trained in a low-resource setting. Those extracted tuples bridge the gap between the free-form dialogue and the compact form of CSKB. Then, a compatibility score is computed to decide how well each extracted tuple aligns with the CSKB.

To train and evaluate ACCENT, we construct the first publicly available event commonsense evaluation dataset for open-domain dialogues (see §4). Besides collecting human commonsense judgments, we request annotators to manually extract event-relation tuples for further analysis.

Our main contributions are three-fold:

- We propose ACCENT, an event commonsense evaluation metric for open-domain dialogue systems. To the best of our knowledge, this is the first work that systematically studies event commonsense in dialogue systems.
- We construct the first publicly available event commonsense evaluation dataset for open-

domain dialogues.<sup>1</sup>

- Extensive experiments show that ACCENT achieves better correlation with human judgments for dialogue commonsense evaluation than several well-designed baselines, and enables easier interpretability of results.

## 2 Background: Event Commonsense

Endowing machines with human-like commonsense reasoning capabilities has been an ultimate goal of artificial intelligence research for decades (McCarthy and Hayes, 1981; LeCun, 2022). While many early works focused on factoid commonsense or the knowledge about concepts (Lenat, 1995; Liu and Singh, 2004), event commonsense emerges as an important aspect for machine commonsense measurement (Chen et al., 2021). Compared with concepts or entities, events are more informative, involving actions, participants, time *etc.* Besides, event commonsense also requires understanding various relations between events (Kuipers, 1984; Rashkin et al., 2018) which would facilitate complex reasoning, especially in interactive scenarios such as dialogues.

Among the current commonsense resources (related works in Appendix A), **ATOMIC<sub>20</sub>** (Hwang et al., 2021) is a comprehensive CSKB including physical-entity, event-centered, and social-interaction knowledge. Its event-centered and social-interaction components take up 84.4% tuples of the entire knowledge base, providing knowledge regarding how events/human actions are associated with other events/actions. For example, given the event “X runs out of stream”, according to **ATOMIC<sub>20</sub>**, this event may happen after “X exercises in the gym”, and the person X is likely to “feel tired”.

## 3 Method

We present ACCENT, as a framework for event commonsense evaluation. Figure 2 gives an overview of ACCENT with two major components.

### 3.1 Symbolic Intermediate Representation

ACCENT uses event-relation tuples as the symbolic intermediate representation. Each tuple contains a head event and a tail event which are connected through an event relation. We formally define events and relations below.

<sup>1</sup>We release ACCENT and our collected datasets at <https://github.com/PlusLabNLP/ACCENT>.

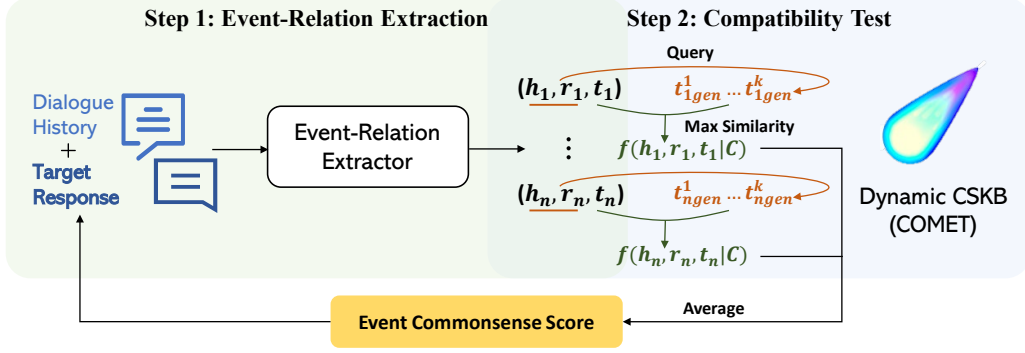


Figure 2: The overview of ACCENT. Given the target response and its dialogue history, ACCENT first extracts the event-relation tuples. Then, the compatibility test (detailed illustration in Figure 4) assigns a score to each tuple: ACCENT queries the dynamic CSKB, *i.e.*, COMET, with  $h$  and  $r$ , and generates  $k$  events. The compatible score is the maximum similarity between the ground-truth  $t$  and the  $k$  generated events  $\{t_{gen}^i\}_{i=1}^k$ . Scores for all tuples in a response are averaged to obtain the event commonsense score for the target response.

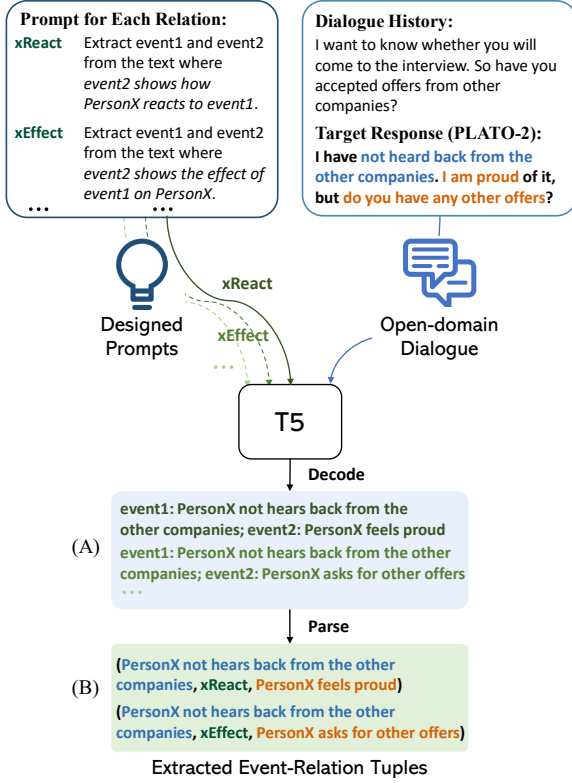


Figure 3: Illustration for event-relation extraction. For each relation  $r \in \tilde{\mathcal{R}}$ , we use its corresponding prompt to guide the model to generate  $h$  and  $t$ . The final tuple is parsed from the generated output.

**Event** Following Pustejovsky et al. (2003), we define events as short phrases with a trigger word and its arguments (*e.g.*, I like to paint). To better align with ATOMIC<sub>20</sub><sup>20</sup>, we normalize the event by replacing tokens referring to people with Person variable (*e.g.*, PersonX likes to paint).

**Relation** We select  $\tilde{\mathcal{R}} = \{xIntent, xWant, oWant, xReact, oReact, xNeed, xAttr, xEffect,$

$oEffect, HinderedBy, IsAfter, HasSubEvent\}^2$  from ATOMIC<sub>20</sub><sup>20</sup> relations. These relations cover human behaviors, *i.e.*, motivation, want, reaction, need, description, towards events (Sap et al., 2019b), the cause-effect and constraint in force dynamic (Talmy, 1988), the temporal information, as well as the parent-child relation in event hierarchy. Examples for each relation are in Appendix C.

### 3.2 Event-Relation Extraction

The input of the event commonsense evaluation task is a list of utterances  $\{u_0, u_1, \dots, u_{n-1}\}$  representing the dialogue history and the target response  $u_n$ . ACCENT first converts the free-form text into event-relation tuples. To retain the information in  $u_n$ , ACCENT extracts tuples whose head and tail events are both from the target response (denoted as “Single”). Besides, to capture event commonsense issues conditioned on the dialogue history (*e.g.*, Figure 1 Example B), ACCENT also extracts tuples whose two events come from  $u_n$  and  $u_{n-1}$  respectively (denoted as “Pair”).

As illustrated in Figure 3, the event-relation extractor in ACCENT is a T5 model  $\mathcal{M}$  (Raffel et al., 2020) guided to generate the head and tail events via designed prompts for each relation. ACCENT concatenates the prompt for  $r \in \tilde{\mathcal{R}}$  and the dialogue as the input and fine-tunes  $\mathcal{M}$  in a low resource setting. When the relation  $r$  exists in the input utterances, the fine-tuned  $\mathcal{M}$  is expected to generate the head and tail events following a particular format, *i.e.*, “event1: {head}; event2: {tail}”, so that the tuple can be parsed from the decoded sequence (from Block A to Block B in Figure 3).

<sup>2</sup>“x” and “o” pertain to PersonX and other person(s).

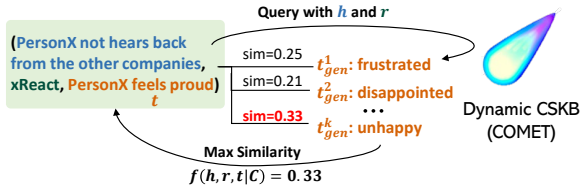


Figure 4: An example of compatibility test. We query the dynamic CSKB with  $h$  and  $r$ , and the compatibility score is the maximum similarity between  $t$  and the generated tail events ( $\{t_{gen}^i\}_{i=1}^k$ ).

Otherwise, the fine-tuned  $\mathcal{M}$  is expected to output “None”. For each  $r \in \tilde{\mathcal{R}}$ , the designed prompt explains the semantic meaning of  $r$  and triggers the model to generate the head and tail events (the prompts are included in Appendix C). At the inference time, we query  $\mathcal{M}$  with prompts for each  $r$  and parse the generated outputs to get  $h$  and  $t$  to construct tuples.

### 3.3 Compatibility Test

After extracting event-relation tuples, ACCENT checks whether these tuples are sensible through a compatibility test. Denoting the CSKB as  $\mathcal{C}$ , the compatibility test aims to learn a scoring function  $f$  based on  $\mathcal{C}$ , where  $f((h, r, t)|\mathcal{C})$  represents the compatibility of the target tuple  $(h, r, t)$  with the CSKB  $\mathcal{C}$ . We propose to score  $(h, r, t)$  by querying a *dynamic version of  $\mathcal{C}$*  with  $h$  and  $r$ . Figure 4 gives an example of the whole process.

Specifically, ACCENT uses COMET (Bosselut et al., 2019) as the dynamic CSKB. COMET adapts the pre-trained language model by fine-tuning it on  $\mathcal{C}$  through a conditional generation task where “{head} {relation} [GEN]” is the source and a tail event is the target. To score  $(h, r, t)$ , we query the model by requiring it to generate  $t_{gen}$  given “{ $h$ } { $r$ } [GEN]”. The beam search method is applied for decoding, so we obtain a set of generated tail events,  $\{t_{gen}^i\}_{i=1}^k$ , where  $k$  is the beam size.

The compatibility score for  $(h, r, t)$  is then computed by checking the similarity between  $t$  and the most similar  $t_{gen}$  among  $\{t_{gen}^i\}_{i=1}^k$ :

$$f((h, r, t)|\mathcal{C}) = \max_{1 \leq i \leq k} \cos(\text{embed}(t), \text{embed}(t_{gen}^i)) \quad (1)$$

Here,  $\text{embed}(\cdot)$  is parameterized by a SentenceBERT model (Reimers and Gurevych, 2019).

After getting the compatibility scores for each extracted tuple, we average them to get the final score for the target response (see Figure 2).

## 4 Data Collection

We construct the first event commonsense evaluation dataset for open-domain dialogues through crowdsourcing on Amazon Mechanical Turk (MTurk). In this section, we describe the collection procedure and the details of the dataset.

### 4.1 Dialogue Data Preparation

We select dialogue histories from DailyDialog (Li et al., 2017), PersonaChat (Zhang et al., 2018), and TopicalChat (Gopalakrishnan et al., 2019) *human-human* dialogues. The dialogue history is limited to at most 4 consecutive utterances. Since human utterances barely contradict event commonsense, to better evaluate machine generated dialogues, we collect responses using advanced dialogue systems, DialoGPT (Zhang et al., 2020d), PLATO-2 (Bao et al., 2021), DiSCoL (Ghazarian et al., 2021).

To ensure most samples contain events and are meaningful for event commonsense evaluation, we filter samples using the following criteria: (1) the response contains at least 5 words; (2) the response contains at least 1 non-interrogative sentence<sup>3</sup>; (3) the response is more than a courtesy (e.g., “It’s been nice chatting with you.”)<sup>4</sup>. After filtering, we randomly select 300 samples and split them into 200 for training and 100 for testing. We name this dataset **DECO** (Dialogue Event Commonsense Dataset).

### 4.2 Tuple Extraction

To train the event-relation extractor of ACCENT, we collect human extracted event-relation tuples from DECO training set. Annotators are shown with the target response, the dialogue history, a specific relation, and are requested to compose event-relation tuples. They are allowed to tick “I cannot find any tuple” if no tuple can be found. We also request them to select whether the tuple belongs to “Single” or “Pair” (defined in §3.2) for each tuple they extract. Figure 8 in Appendix D shows our data collection panel. We launched HITs<sup>5</sup> for relations in  $\tilde{\mathcal{R}}$  repeatedly until we obtained at least 20 tuples for each relation. In order to ensure the test set is comprehensive, we particularly request annotators to compose tuples for all 12 relations in  $\tilde{\mathcal{R}}$  (100 samples  $\times$  12 relations in total).

<sup>3</sup>We check this by finding sentences that are not ended with a question mark (“?”).

<sup>4</sup>These responses are manually filtered out.

<sup>5</sup>HIT is an assignment unit on Amazon MTurk.

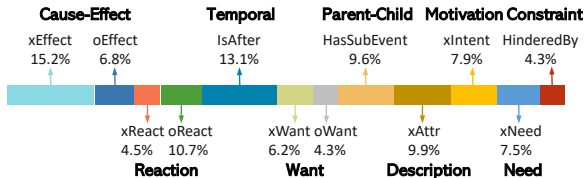


Figure 5: Relation distribution in DECO test set.

A separate validation round was conducted to check whether each extracted tuple satisfies (1) the head and tail are events, (2) the head and tail come from  $u_n$  or  $u_{n-1}$ , (3) the relation between the head and tail can be inferred from the dialogue. A tuple is deemed valid if the majority of 3 annotators vote “yes”. After removing invalid tuples (the dialogue numbers remain unchanged), we collected 307 tuples for training and 467 tuples from the DECO test set. Figure 5 shows the relation distribution in the densely annotated test set. More details about DECO statistics are included in Appendix D.

### 4.3 Commonsense Scoring

We instruct annotators to score target responses in terms of event commonsense by focusing on the events and their relations (the guideline is shown in Figure 7). Each response was annotated by 3 individual annotators with a scale of 1 to 5. Following Mehri and Eskenazi (2020), we measure the inter annotator agreement (IAA) by correlating each annotation with the mean of the other annotations for the same sample, and the Spearman correlation is 0.578 showing an acceptable agreement.<sup>6</sup> The final event commonsense score assigned to each sample is the average of 3 individual ratings.

We also requested the annotators to judge the overall commonsense of a dialogue response before introducing event commonsense to annotators. Among the 900 annotation pairs we collected, the Spearman correlation between the two scores reaches 0.862, which indicates that *event commonsense is a key component in overall commonsense reasoning*.

### 4.4 Additional Human-Machine Dialogues

We further explore the generalization ability of ACCENT on responses with *human-machine* dialogue histories. We select 100 samples from ConTurE (Ghazarian et al., 2022a), a turn-level evaluation dataset, to annotate event commonsense scores. We denote this dataset as **ConTurE Subset**. Its statistics are also included in Appendix D.

<sup>6</sup>0.40-0.69 implies strong relationship.

## 5 Experiments

### 5.1 Setups

We compare ACCENT with baseline methods for event commonsense evaluation and also examine its two components separately. Therefore, our experiments include three setups for the evaluation: **Setup 1 (Metrics Performance)** Our main goal is to evaluate the commonsense metric, and we achieve this by computing the correlation between automatic scores and human judgments. ACCENT and baseline metrics are tested on DECO test set and ConTurE Subset.

**Setup 2 (Event-Relation Extraction)** We evaluate the performance of the event-relation extraction component of ACCENT by comparing the automatically extracted tuples with human extracted tuples on DECO test set. We view checking whether a tuple with relation  $r$  is extracted from the utterances  $u_n$  and  $u_{n-1}$  as a binary classification problem and compute the F1 score. We also measure how “close” the automatically extracted head and tail events are to human extraction results. We convert the tuple into a sentence by concatenating the head and tail, and then compute BLEU-2 (Papineni et al., 2002) and BERTScore (Zhang et al., 2020c).

**Setup 3 (Compatibility Test):** The compatibility test component of ACCENT can be viewed as a tuple scoring task. We compare our proposed approach with other tuple scoring methods on a large-scale benchmark (Fang et al., 2021a) which contains event-relation tuples with 0 (compatible to a given CSKB) or 1 (not compatible to the CSKB) scores. Since the training relations in this benchmark differ from relations supported by the off-the-shelf COMET, we train our own COMET on its training set (see Appendix E.2 for more details) to make our compatibility test component applicable to this test set. This benchmark dataset covers all 12 relations in  $\tilde{\mathcal{R}}$  as well as 6 more relations.

### 5.2 Baselines

We compare ACCENT with 5 baseline metrics: (1, 2) **FED-understandable/appropriate** (Mehri and Eskenazi, 2020) are two off-the-shelf baselines. “Understandable” and “Semantically Appropriate” are closer to commonsense compared to the rest of the criteria in FED. (3) **Cross-encoder** is a widely used model for sentence-pair regression tasks. We use BART (Lewis et al., 2020) as the backbone. (4) **Cross-encoder (COMET)** is a variant of (3) with COMET trained on ATOMIC<sub>20</sub><sup>20</sup> as the back-

bone. (5) **MLP regressor** (Zhou et al., 2021) is trained with neural features from DialoGPT and symbolic features from ConceptNet (details in §7). The cross-encoders and the MLP regressor require event commonsense scores to train the model in an end-to-end manner. We use the annotated scores in DECO training set to train them, and split 20% data for validation to conduct hyperparameter search.

For Setup 2, we consider the following baseline approaches: (1) **ASER Extractor** (Zhang et al., 2020b) first extracts events through patterns from dependency parsing and then uses a neural classifier to predict the relation. (2) **CSKB Search** (Zhou et al., 2021) searches the one-hop neighbors in ATOMIC<sub>20</sub> through keyword matching.

For Setup 3, we consider 4 tuple scoring baselines. These baselines convert a tuple to an embedding and train a binary classifier to give score: (1) **BERT** feeds  $h, r, t$  to BERT and concatenates their [CLS] embeddings to get the tuple embedding. (2) **BERTSAGE** (Fang et al., 2021b) further concatenates the average embedding of the neighbors of  $h$  and  $t$  in an event knowledge graph. (3) **KG-BERT** (Yao et al., 2019) inputs “[CLS],  $h$ , [SEP],  $r$ , [SEP],  $t$ ” to get the tuple embedding. (4) **KG-BERTSAGE** (Fang et al., 2021a) further concatenates the average embedding of neighboring nodes. We use ROBERTa<sub>LARGE</sub> (Liu et al., 2020) as the backbone which has roughly the same parameter budget with COMET to have a fair comparison.

The details of the baseline implementations are in Appendix E.1.

### 5.3 ACCENT Implementation

The proposed ACCENT framework is implemented using the Transformers library (Wolf et al., 2020). For event-relation extraction, we fine-tune T5-base<sup>7</sup> for 50 epochs with the batch size of 4 and the learning rate of 5e-5. The training data comes from the human extracted tuples from DECO training set. We additionally select 5 negative samples (dialogues that do not have a certain relation) per relation from the training set and set their target output as “None” to guide the model to handle cases which do not contain a certain relation. During inference, if no tuple is extracted after considering all relations, we assign a score of 0.5 to the sample. For compatibility test, we use the off-the-shelf COMET model trained on

<sup>7</sup>[huggingface.co/t5-base](https://huggingface.co/t5-base)

	DECO		ConTurE	
	$\gamma$	$\rho$	$\gamma$	$\rho$
FED-appropriate	-0.16	-0.10	-0.09	-0.04
FED-understandable	-0.12	-0.07	-0.08	-0.04
Cross-encoder	0.15	0.15	-0.05	-0.09
Cross-encoder (COMET)	0.17	0.17	0.00	0.00
MLP Regressor	0.11	0.01	0.17	0.16
ACCENT (Ours)	<b>0.30</b>	<b>0.30</b>	<b>0.21</b>	<b>0.22</b>

Table 1: Pearson ( $\gamma$ ) and Spearman ( $\rho$ ) correlations between human judgments and different automatic evaluation metrics. The results for ACCENT are all significant ( $p < 0.05$ ).

ATOMIC<sub>20</sub> (Hwang et al., 2021)<sup>8</sup>. When querying COMET through generation, we use beam search with a beam size of 10 to get commonly sensible tail events. The embed( $\cdot$ ) in Equation (1) is parameterized by `paraphrase-MiniLM-L6-v2` provided in the Sentence-Transformers library<sup>9</sup>.

## 6 Results and Analysis

### 6.1 Metrics Performance

Table 1 shows the correlations between automatic scores and human annotations. ACCENT uniformly outperforms the baselines on both two test sets. Specifically, off-the-shelf metrics (“FED-appropriate”, “FED-understandable”) perform poorly. For “Cross-encoder (COMET)”, its results show that implicitly using the CSKB through transfer learning cannot yield satisfactory performance. Besides, cross-encoders fail to generalize well to ConTurE Subset whose dialogue histories are from human-machine dialogues. For “MLP Regressor”, although it tries to utilize the CSKB explicitly, it is not as effective as ACCENT.

Some examples from the DECO test set and their event commonsense scores given by ACCENT are shown in Table 2. These scores are close to human judgements and enjoy great interpretability owing to the extracted event-relation tuples.

Apart from the sample-level correlation, we further examine whether ACCENT can reflect model performance in terms of event commonsense. Figure 6 shows the rankings of three dialogue systems used in DECO construction given by human and ACCENT. Human and ACCENT rank the three systems exactly the same and the two sets of averaged scores highly correlates with each other.

<sup>8</sup>[github.com/allenai/comet-atomic-2020](https://github.com/allenai/comet-atomic-2020)

<sup>9</sup>[huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2](https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2)

Dialogue History	Target Response	Human	ACCENT	Extracted Tuple
Yes, I run them all the time! How exciting, are you ready?	I'm ready for the marathon! So I'm going to be running a half marathon in a few weeks.	0.80	0.61	(PersonX is ready for the marathon, xEffect, PersonX run a half marathon)
I like pizza as long as it's a meat lover. I hate vegetables.	Lol, yes it is! Great for you to eat so healthy!	0.53	0.45	(PersonX hate vegetables, xAttr, PersonX be healthy)
I had an accident.	That's what I like to call the most beautiful thing.	0.27	0.23	(PersonX has an accident, oWant, PersonY wants to call the most beautiful thing)

Table 2: Examples of evaluation scores given by ACCENT from the DECO test set. Only  $u_{n-1}$  in the dialogue history is shown, and the human annotations are normalized to the range of (0, 1) for better demonstration. ‘‘Extracted Tuple’’ column shows tuples extracted in ACCENT framework which account for the given scores.

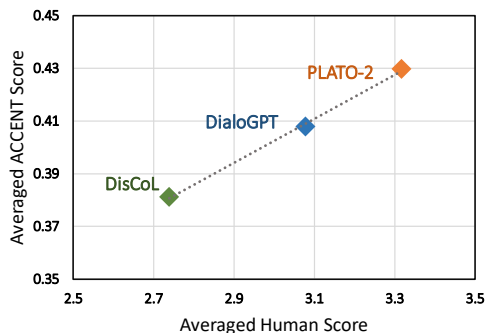


Figure 6: Average event commonsense scores of generated responses of different models using human annotations (scale 1 to 5) and ACCENT automatic evaluation (scale 0 to 1). The rankings of systems given by human and ACCENT are the same.

	P	R	F1	BLEU	BERTScore
CSKB Search	29.9	<b>96.3</b>	<b>45.7</b>	26.9	89.8
ASER Extractor	<b>31.5</b>	23.6	27.0	32.4	89.3
<b>Ours</b>	31.4	55.0	40.0	<b>41.6</b>	<b>93.5</b>

Table 3: Performances of different event-relation extraction methods on DECO test set. P: Precision. R: Recall.

## 6.2 Tuple Extraction Performance

Table 3 shows the results of Setup 2 where we evaluate the event-relation extraction performance on DECO test set. Our proposed method achieves much higher BLEU and BERTScore than two baselines, indicating that the composed events in tuples have reasonable quality. However, joint event-relation extraction remains challenging because it combines the event extraction and relation identification. Although our proposed method has higher score than ASER Extractor by F1, it still has plenty of room for improvement. As for CSKB Search, it usually returns a lot of tuples, thus resulting in high recall and very poor precision. Also, searching CSKB is not applicable in our framework because this method can only return sensible tuples.

	Subset	All
BERT	62.0±0.3	61.5±0.3
BERTSAGE	55.8±0.7	55.8±0.7
KG-BERT	62.6±0.7	62.3±0.8
KG-BERTSAGE	63.2±0.4	62.9±0.3
<b>Ours</b>	<b>68.0±0.8</b>	<b>67.6±0.8</b>

Table 4: Test results on the CSKB compatibility benchmark. We report the overall AUC across all relations (‘‘All’’) and the AUC across samples with our target relations (‘‘Subset’’). Both the averaged metric and its standard deviation are reported over 3 runs.

		DECO	ConTurE
<b>ACCENT (whole)</b>		<b>0.30</b>	<b>0.22</b>
I	ASER Extractor	0.14	0.00
	w/o Pair	0.19	0.08
	w/o Single	0.24	0.18
	Gold Tuples	<b>0.42</b>	-
II	Bert	-0.08	0.09
	KG-Bert	0.13	0.19
	COMET (neural)	0.16	0.05

Table 5: Ablation results measured by Spearman correlation. I: Ablation of the event-relation extraction part. The gray row shows the results using human extracted tuples which provides an upper bound. II: Ablation of the compatibility test part of ACCENT.

## 6.3 Compatibility Test Performance

Table 4 depicts the test results on the benchmark dataset. Our method outperforms all baselines, and it does not require negative samples for training. The major difference between our method and those tuple scoring baselines is that we use the tuples in the existing CSKB to train a dynamic CSKB, *i.e.*, COMET, instead of a discriminative model. We assume our strong results may be due to the generalization ability of the dynamic CSKB.

## 6.4 Ablation Studies

We conduct ablation studies to explore (1) whether the proposed event-relation extraction method can

	STS Avg.	DECO	ConTurE
Sentence-BERT	79.82	0.30	0.22
DiffCSE <sup>10</sup>	78.21	0.12	0.25
ESimCSE <sup>11</sup>	77.44	0.19	0.24
Sup-SimCSE <sup>12</sup>	<b>82.52</b>	<b>0.31</b>	<b>0.26</b>

Table 6: Results with different sentence embedding methods measured by Spearman correlation. Following Gao et al. (2021), we use the average results on the semantic textual similarity (STS) tasks to reflect the sentence embedding performance.

lead to better final metric performance; (2) given the automatically extracted tuples, whether the proposed compatibility test method can lead to higher correlation with human judgment.

To answer (1), we compare different methods to get the event-relation tuples (Part I in Table 5). Among the event-relation extraction baselines, we only consider ASER Extractor because CSKB search is not applicable in our framework as discussed in §6.2. Note that the event-relation extractor in ACCENT considers tuples in both “Single” and “Pair” settings to cover two potential types of errors (see §3.2). To verify this, we compare the variations of our proposed method where we only use tuples marked as “Single” or “Pair” for model training. Also, the human extracted tuples in DECO test set are used to provide an upper bound.

To answer (2), we fix the event-relation extraction part and change the compatibility test part (Part II in Table 5). We consider **BERT** and **KG-BERT** trained on the CSKB compatibility benchmark because they do not need event graph information and can be seamlessly applied to our compatibility test. Also, while we query COMET through tail generation, another intuitive design is using the model loss with “{h} {r} [GEN]” as the source and  $t$  as the target to give scores. We map the loss to  $(0, 1)$  through an exponential function, and name this alternative as “**COMET (neural)**” for it skips the symbolic decoding of  $t_{gen}$ .

Table 5 demonstrates that the whole ACCENT gives the best result. Considering the variations of our design, “w/o Pair” gives much lower results, indicating that limiting the symbolic intermediate representation to only the information contained in the target response is not enough. This observation is in accord with our finding that some event commonsense errors occur when we take the dialogue history into account.

Another empirical discovery is that although “COMET (neural)” is a direct way of using the dynamic CSKB, its performance is poorer than what we propose in ACCENT. We assume that comparing  $t$  and  $t_{gen}$  in a symbolic fashion can yield more comparable scores among tuples with different relations (details in Appendix F).

In our implementation of ACCENT, the comparison of  $t$  and  $t_{gen}$  is done by calculating the cosine similarity between their Sentence-BERT embeddings. We further experiment with other sentence embedding methods based on contrastive learning. Specifically, we consider DiffCSE (Chuang et al., 2022), ESImCSE (Wu et al., 2022) which are two unsupervised contrastive learning frameworks for learning sentence embeddings. We also consider Sup-SimCSE (Gao et al., 2021) which leverages annotated natural language inference datasets by using “entailment” pairs as positives and “contradiction” pairs as hard negatives in the contrastive learning objective. As shown in Table 6, ACCENT can benefit from the improvement of the sentence embedding method, *i.e.*, using Sup-SimCSE (Gao et al., 2021). We support both Sentence-BERT and Sup-SimCSE in our released ACCENT codebase.

## 6.5 Error Analysis

Since ACCENT is a pipeline framework, there is likely error propagation. In section 6.4, we rule out the errors introduced by the event-relation extraction component by using human-extracted gold tuples. Results show that ACCENT with gold tuples (see “Gold Tuples” in Table 5) gives higher correlation with human judgment than “ACCENT (whole)” which uses the model-extracted tuples, indicating that ACCENT can benefit from high quality symbolic intermediate representation. We further include a qualitative analysis of the automatically extracted tuples in Appendix G, and believe improving the joint event-relation extraction is a worthwhile direction for future work.

## 7 Related Work

### Automatic Open-Domain Dialogue Evaluation

The evaluation of open-domain dialogue systems has long been a challenge due to the system’s open-

<sup>10</sup><https://huggingface.co/voidism/diffcse-roberta-base-sts>

<sup>11</sup><https://huggingface.co/ffgcc/esimcse-roberta-base>

<sup>12</sup><https://huggingface.co/princeton-nlp/sup-simcse-roberta-base>



ended goal (Huang et al., 2020), and simply scoring the overall quality is far from enough (Finch and Choi, 2020). Thus, researchers have decomposed the evaluation of open-domain dialogues into multiple facets and developed corresponding automatic evaluation metrics (Pang et al., 2020; Mehri and Eskenazi, 2020). While aspects like context coherence (Tao et al., 2018; Ghazarian et al., 2022b), diversity (Hashimoto et al., 2019), engagement (Ghazarian et al., 2020), have been systematically studied in the literature, the aspect of commonsense has long been neglected.

The closest related work is Zhou et al. (2021) which is mainly about commonsense-focused dialogues collection but also proposes an automatic metric for commonsense evaluation by training an MLP regressor on both symbolic and neural features. The symbolic features include the numbers of one-hop and two-hop triplets in ConceptNet (Speer et al., 2017) that can be found between the target response and its dialogue history. Although this metric utilizes the CSKB explicitly, it is limited to the direct search with surface form and only considers the number of triplets, and the CSKB used in the work is more about concepts but not event commonsense.

**Joint Event-Relation Extraction** While event extraction (Ahn, 2006) and relation identification (Do et al., 2011) are well-studied, how to jointly acquire them remains a challenge. We argue that joint event-relation extraction is an important problem because in practical use cases, the input is usually free-form text without pre-extracted events. Zhang et al. (2020b) is a pioneer work trying to jointly extract event and relation through a pipeline to automatically construct large knowledge graphs. Researchers in this work resort to rule-based methods for event extraction and train a classifier to predict the relation between a pair of events.

**CSKB Compatibility** CSKB population enlarges CSKB automatically by adding new links or nodes which are compatible with the commonsense knowledge to the existing CSKB. In Fang et al. (2021a,b), researchers try to add events from a large event knowledge graph to a CSKB. Compatibility test component of ACCENT is relevant to CSKB population task and it is defined in a more general setting where the head and tail of the tuple can be arbitrary events.

## 8 Conclusion

We present ACCENT, an automatic evaluation metric for event commonsense evaluation of open-domain dialogue systems. We show that by using event-relation tuples as the symbolic intermediate representations, ACCENT can effectively utilize the CSKB and achieve a decent correlation with human judgments for dialogue commonsense evaluation.

## 9 Limitations

In this work, we conduct research on event commonsense of open-domain dialogue systems for the first time. While achieving higher correlations with human judgments than existing baselines, ACCENT has some limitations:

First, the ACCENT framework is based on a fixed set of event relations and the commonsense knowledge in ATOMIC<sub>20</sub><sup>20</sup> which may fail to cover some potential event commonsense aspects. We believe augmenting the current framework with more commonsense resources is a worthwhile direction for the further improvement of ACCENT.

Second, the event-relation extractor in ACCENT framework is a T5 model fine-tuned in a low resource setting. Although the current model can yield fairly strong performance, it is an important research direction to improve the joint event-relation extraction component because the extracted tuples serve as the symbolic representation for commonsense reasoning in ACCENT framework. Since human extracted tuples are very costly to collect, we hope to explore whether we can improve this component through high-quality synthetic data construction or transfer learning in the future.

## 10 Acknowledgments

We thank the PlusLab members and the anonymous reviewers for their constructive feedback. This work is supported in part by the DARPA Machine Common Sense (MCS) program under Cooperative Agreement N66001-19-2-4032, and a Meta Sponsored research award.

## References

David Ahn. 2006. *The stages of event extraction*. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

- Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. [PLATO-2: Towards building an open-domain chatbot via curriculum learning](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. [COMET: Commonsense transformers for automatic knowledge graph construction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.
- Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.
- Muhao Chen, Hongming Zhang, Qiang Ning, Manling Li, Heng Ji, Kathleen McKeown, and Dan Roth. 2021. [Event-centric natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 6–14, Online. Association for Computational Linguistics.
- Yung-Sung Chuang, Rumén Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shangwen Li, Scott Yih, Yoon Kim, and James Glass. 2022. [DiffCSE: Difference-based contrastive learning for sentence embeddings](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.
- Quang Do, Yee Seng Chan, and Dan Roth. 2011. [Minimally supervised event causality identification](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Nouha Dziri, Ehsan Kamaloo, Kory W Mathewson, and Osmar Zaiane. 2018. Augmenting neural response generation with context-aware topical attention. *arXiv preprint arXiv:1811.01063*.
- Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. [Benchmarking commonsense knowledge base population with an effective evaluation dataset](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8949–8964, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. [Discos: Bridging the gap between discourse knowledge and commonsense knowledge](#). In *Proceedings of the Web Conference 2021*, pages 2648–2659.
- Sarah E. Finch and Jinho D. Choi. 2020. [Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Brett M Frischmann. 2021. Common sense commons: The case of commonsensical social norms. *Available at SSRN 3781955*.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sarik Ghazarian, Behnam Hedayatnia, Alexandros Pangelis, Yang Liu, and Dilek Hakkani-Tur. 2022a. [What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4194–4204, Dublin, Ireland. Association for Computational Linguistics.
- Sarik Ghazarian, Zixi Liu, Tuhin Chakrabarty, Xuezhe Ma, Aram Galstyan, and Nanyun Peng. 2021. [DiS-CoL: Toward engaging dialogue systems through conversational line guided response generation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 26–34, Online. Association for Computational Linguistics.
- Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796.
- Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022b. [DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.
- Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. [Topical-Chat: Towards Knowledge-Grounded](#)

- Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.
- Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. **Unifying human and statistical evaluation for natural language generation**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.
- Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.
- 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 *AAAI*.
- Benjamin Kuipers. 1984. Commonsense reasoning about causality: deriving behavior from structure. *Artificial intelligence*, 24(1-3):169–203.
- Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27.
- Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. **BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. **DailyDialog: A manually labelled multi-turn dialogue dataset**. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. **Ro{bert}a: A robustly optimized {bert} pretraining approach**.
- John McCarthy and Patrick J Hayes. 1981. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier.
- Shikib Mehri and Maxine Eskenazi. 2020. **Unsupervised evaluation of interactive dialog with DialoGPT**. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.
- Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. **GLUCOSE: Generalized and Contextualized story explanations**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.
- Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. **Towards holistic and automatic evaluation of open-domain dialogue generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. 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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. **Language models as knowledge bases?** In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.
- 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**. *Journal of Machine Learning Research*, 21(140):1–67.
- Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. **Event2Mind: Commonsense inference on events, intents, and reactions**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

- Long Papers*), pages 463–473, Melbourne, Australia. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. **Sentence-BERT: Sentence embeddings using Siamese BERT-networks**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. **Atomic: An atlas of machine commonsense for if-then reasoning**. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. **Social IQa: Commonsense reasoning about social interactions**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. **Commonsense reasoning for natural language processing**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. **Conceptnet 5.5: An open multilingual graph of general knowledge**. In *Thirty-first AAAI conference on artificial intelligence*.
- Leonard Talmy. 1988. Force dynamics in language and cognition. *Cognitive science*, 12(1):49–100.
- Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. **Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems**. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. **Transformers: State-of-the-art natural language processing**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. **ESim-CSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding**. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3898–3907, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. **Kgbert: Bert for knowledge graph completion**. *arXiv preprint arXiv:1909.03193*.
- Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020a. **Transomcs: From linguistic graphs to commonsense knowledge**. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4004–4010. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020b. **ASER: A large-scale eventuality knowledge graph**. In *WWW*, pages 201–211.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. **Personalizing dialogue agents: I have a dog, do you have pets too?** In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020c. **Bertscore: Evaluating text generation with bert**. In *International Conference on Learning Representations*.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020d. **DIALOGPT : Large-scale generative pre-training for conversational response generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.
- Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. **Commonsense-focused dialogues for response generation: An empirical study**. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132, Singapore and Online. Association for Computational Linguistics.

## A Commonsense Knowledge Bases

To endow machines with commonsense reasoning abilities, a growing number of CSKBs are developed through human annotation and information extraction. From earlier CSKBs, ConceptNet (Liu and Singh, 2004; Speer et al., 2017) focuses more on taxonomic (*e.g.*, “IsA”) and lexical (*e.g.*, “Synonym”, “RelatedTo”) knowledge; TransOMCS (Zhang et al., 2020a) automates the knowledge base construction by leveraging the same limited set of relations defined in ConceptNet.

Recent CSKBs give more focus on event commonsense knowledge. In this work, we select ATOMIC<sub>20</sub><sup>20</sup> (Hwang et al., 2021) as the knowledge source of ACCENT framework because it is a comprehensive CSKB with rich knowledge regarding how events and human actions are associated with each other. For comparison, ATOMIC (Sap et al., 2019a), as the pioneer of ATOMIC<sub>20</sub><sup>20</sup>, consists of only nine relations and therefore poses limitations. Another recent event-centric CSKB is GLUCOSE (Mostafazadeh et al., 2020), which however focuses on a specific part of event commonsense (mostly on causal inference) and is less comprehensive and suitable for our work.

## B Pseudo Code of ACCENT

In §3, we introduce the symbolic intermediate representation and the two components in ACCENT. Algorithm 1 displays the skeleton of the whole framework.

Line 3-9 in Algorithm 1 show the joint event-relation extraction in ACCENT. We query the event-relation extraction model  $\mathcal{M}$  with prompts for each relation. The head and tail events can be parsed from the generated result if it is not “None” and follows the pre-defined format, *i.e.*, “event1: {head}; event2: {tail}”. Line 16-21 in Algorithm 1 show the compatibility test in ACCENT. Each tuple is given a score based on the maximum cosine similarity between its tail and the commonsense tails obtained from the dynamic CSKB  $\mathcal{C}$ . After calculating scores for each extracted tuple, we average them to get the event commonsense score for the target utterance (Line 24 in Algorithm 1).

## C Event Relations

As introduced in §3.1, ACCENT selects relations from ATOMIC<sub>20</sub><sup>20</sup> which are related to event commonsense. These event relations can help cover

---

### Algorithm 1: ACCENT framework.

---

**Input** : Dialogue history  $h$ , target utterance  $u$ , prompt dict  $P$ , extractor  $\mathcal{M}$ , dynamic CSKB  $\mathcal{C}$ , sentence embedder  $\mathcal{E}$

**Output** : Event commonsense score  $s$

```
1 tuples  $\leftarrow$  []
2 // Event-relation extraction.
3 foreach (rel, prompt) in  $\mathcal{P}$  do
4   raw_output  $\leftarrow$  generate( $\mathcal{M}$ ,  $h$ ,  $u$ ,  $p$ )
5   if check_format(raw_output) then
6     (head, tail)  $\leftarrow$  parse(raw_output)
7     tuples.append((head, rel, tail))
8   end
9 end
10 // Compatibility test.
11 if is_empty(tuples) then
12   return 0.5
13 else
14   tuple_scores  $\leftarrow$  []
15   foreach (head, rel, tail) in tuples do
16     score  $\leftarrow$  0
17     cs_tails  $\leftarrow$  query( $\mathcal{C}$ , head, rel)
18     foreach cs_tail in cs_tails do
19        $x \leftarrow \cos(\mathcal{E}(\text{tail}), \mathcal{E}(\text{cs\_tail}))$ 
20       score  $\leftarrow \max(\text{score}, x)$ 
21     end
22     tuple_scores.append(score)
23   end
24   return average(tuple_scores)
25 end
```

---

different types of insensibility for event commonsense evaluation of open-domain dialogues. Table 7 shows examples in DECO where the system response violates event commonsense in terms of different types of event relations.

Note that although “Cause” and “xReason” in ATOMIC<sub>20</sub><sup>20</sup> are also related to event commonsense, we exclude them from the selected subset  $\tilde{\mathcal{R}}$ . This is because the cause-effect relation can be covered by “xEffect”/“oEffect” and tuples with “Cause” and “xReason” relations take up less than 0.1% percent of ATOMIC<sub>20</sub><sup>20</sup>. Moreover, we exclude “IsBefore” because a tuple with “IsBefore” relation can be equivalently converted to a tuple with “IsAfter” relation by switching the head and tail. As shown in Table 8, for each relation in  $\tilde{\mathcal{R}}$ , a prompt is manually designed to explain its semantic meaning. These designed prompts give more hints to the pre-trained model and allow a single model to extract tuples for different relations.

## D Additional Details of Data Collection

### D.1 Quality Control

To ensure the annotators have a good understanding of event and event commonsense, we restrict the annotators from English-speaking countries, and those who have finished at least 5,000 HITs with an acceptance rate  $> 97\%$ . The compensation rate for annotators is calculated using a per hour wage of \$16.<sup>13</sup>

For commonsense scoring (see §4.3), we requested 3 annotators to score each sample, and we instructed them to specifically consider events and their relations in the dialogue to give the event commonsense score. Figure 7 shows the annotation guideline we used for event commonsense scoring. We also set a sample for attention check in each HIT. HITs that failed the check were reassigned to other annotators.

For tuple extraction (see §4.2), we conducted a training round before the large scale annotation and 8 annotators proceeded to the final round. Each HIT in this task was assigned to 2 individual annotators. The template used to collect event-relation tuples is shown in Figure 8. When validating the extracted tuples, 3 annotators judged each tuple, and we achieved Fleiss’ Kappa (Fleiss, 1971)  $\kappa = 0.491$  (moderate agreement). Tuples marked

<sup>13</sup>We pay \$1 per HIT for the scoring task and \$3 per HIT for the tuple extraction task. An additional bonus is sent to annotators who successfully pass the training round.

as invalid by the majority vote are not included in the final dataset.

### D.2 Dataset Statistics

Table 9 gives the statistics of DECO and ConTurE Subset. Although machine generated responses in DECO are given by advanced open-domain dialogue systems, some event commonsense errors still exist. For ConTurE Subset, we use it to test the generalization ability of different metrics. Table 10 gives the numbers of human extracted event-relation tuples. Note that the test set of DECO is thoroughly annotated (we consider every relation on each sample) to provide a test set for the joint event-relation extraction task. All the data we collected are in English.

### D.3 License

The development of DECO and ConTurE Subset is based on the dialogues coming from DailyDialog, PersonaChat, TopicalChat, and ConTurE. PersonaChat<sup>14</sup>, TopicalChat<sup>15</sup> and ConTurE<sup>16</sup> are licensed. We ensure that we did not violate any license conditions when developing our datasets.

## E Additional Details of Experiment

This section describes more details about baseline implementation, applying ACCENT to CSKB benchmark, and computational resources. The implementation details of the proposed ACCENT framework are discussed in §5.3

### E.1 Baseline Implementation

We compare ACCENT with 5 baseline metrics on event commonsense evaluation. All the metrics are tested on DECO test set and ConTurE Subset. For metrics which require training, the training set of DECO is used, and we split 20% data for validation. The implementation details are as follows:

- **FED-understandable/appropriate:** We use their released model<sup>17</sup>.
- **Cross-encoder:** We use the cross-encoder with a regression head implemented in the Sentence-Transformers library (Reimers and

<sup>14</sup><https://github.com/facebookresearch/ParlAI/blob/main/LICENSE>

<sup>15</sup><https://github.com/alexa/Topical-Chat/blob/master/DATALICENSE>

<sup>16</sup><https://github.com/alexa/conture/blob/main/DATALICENSE>

<sup>17</sup><https://github.com/Shikib/fed>














Relation	Negative Example	Event-Relation Tuple
xIntent Motivation	 A: Stay around for a while. The party is warming up. B: We'll need to <b>get you some ice cream</b> , you know, <b>to warm up your body</b> .	(PersonX gets PersonY some ice cream, xIntent, PersonX warms up PersonY's body)
xNeed Need	 A: I boated around the world with my husband when we were younger. B: I love boating. I also like to paint , I just <b>need an oxygen mask</b> . I need a life.	(PersonX loves boating, xNeed, PersonX needs an oxygen mask), (PersonX likes to paint, xNeed, PersonX needs an oxygen mask)
xReact, oReact Reaction	 A: That is funny! At work they make me wear a uniform, boohoo! B: That is unfortunate, I actually lost my arm from a car accident so I do not have to.  A: <b>That is interesting</b> ! Do you know Apple has anything to say on that?	(PersonX loses PersonX's arm from a car accident, oReact, PersonY feels interesting)
xWant, oWant Want	 A: We don't give bonus every month, but we offer a semi-annual bonus. And you will receive two weeks paid vacation a year, as well. Does it suit you? B: Yes, thank you. <b>May I ask for an apartment</b> ?  A: No... <b>I want to take your word on that one</b> ! It'll be all I need :)	(PersonX asks for an apartment, oWant, PersonY wants to take PersonX's word)
xAttr Description	 A: Are you a vegetarian? I am. B: Yes I am. I do not like meat.  A: <b>I'm a vegetarian</b> and I love meat .	(PersonX loves meat, xAttr, PersonX be a vegetarian)
xEffect, oEffect Cause-Effect	 A: How you celebrate your Valentine's Day with your wife? B: I am not sure about you, but <b>my wife is not into Valentine's day</b> ... So <b>we celebrate a lot</b> .	(PersonX be not into Valentine's day, xEffect, PersonX celebrates a lot)
HinderedBy Constraint	 A: My mom does not bake, she does not even cook. B: My mom used to cook for my family, but I think <b>my mom's got too big</b> to cook anything for anymore.	(PersonX cooks for family, HinderedBy, PersonX gets too big)
IsAfter Temporal	 A: Marco has fallen off a ladder. I think he's hurt his back. What shall we do? B: <b>Marco is still on the ladder</b> , it just got knocked over . Marco will not get any sleep.	(PersonX be on the ladder, isAfter, PersonX gets knocked over)
HasSubEvent Parent-Child	 A: Yeah he was an internal medicine practitioner before he turned to comedy so <b>he attended to the woman</b> until medics arrived. B: Ohhh I see. I thought he was in the audience when <b>he was having the seizure</b> .	(PersonX attends to the woman, HasSubEvent, PersonX has the seizure)

Table 7: Event relations with corresponding negative examples in DECO.  denotes responses generated by open-domain dialogue systems. Each example contains events (highlighted with green and yellow) which violate event commonsense in terms of the corresponding event relation. Such event commonsense errors can be captured by nonsensical event-relation tuples.

Relation	Semantic Meaning	Designed Prompt (Extract event1 and event2 from the text where ...)
xIntent	because PersonX wanted	event2 shows PersonX's intent for event1.
xNeed	but before, PersonX needed	event2 needs to be true for event1 to take place.
xReact	as a result, PersonX feels	event2 shows how PersonX reacts to event1.
oReact	as a result, Y or others feels	event2 shows how PersonY reacts to event1.
xWant	as a result, PersonX wants	event2 shows what PersonX wants after event1 happens.
oWant	as a result, Y or others wants	event2 shows what PersonY wants after event1 happens.
xAttr	X is seen as	event2 shows how PersonX is viewed as after event1.
xEffect	as a result, PersonX will	event2 shows the effect of event1 on PersonX.
oEffect	as a result, Y or others will	event2 shows the effect of event1 on PersonY.
HinderedBy	can be hindered by	event1 fails to happen because event2.
IsAfter	happens after	event1 happens after event2.
HasSubEvent	includes the event/action	event1 includes event2.

Table 8: Semantic meanings and designed prompts for the selected ATOMIC<sub>20</sub><sup>20</sup> relations. The semantic meanings are from Hwang et al. (2021).

In this survey, we are specifically interested in the **event commonsense**. The **core of event commonsense concerns the events and the relations between them**. So you need to focus on the events in the **response** or its history and whether they are sensibly related to each other.

**Hint** More explanation for **Event** and **Relation**:

**Event**

An **event** is something that happens (e.g., actions) or processes a certain state (e.g., properties). We can usually describe an **event** through a verb-centric phrase. For example, in "I have participated in the 3000-meter race so I feel so tired now.", we can find two events, **I participate in the 3000-meter race** and **I feel tired**.

**Relation**

**Relation** is about how **one event** relates to **another event**. In the former example, there exists **a relation of "Effect" - "I feel tired"** is the **effect** of **"I participate in the 3000-meter race"**.

Common **relations** in event commonsense are

- **Motivation**: A person does **one event** because he / she hopes to achieve **another event**.
- **Prerequisite**: A person needs **one event** to make **another event** happen.
- **Constraint**: **One event** may fail to happen because of **another event**.
- **Effect**: **One event** be the cause and **another event** be the effect.
- **Reaction**: **One event** (usually about the feeling or emotion) is the reaction to **another event**.
- **Desire**: **One event** happens, which triggers a person to want **another event** to happen.
- **Description**: **One event** can be described as **another event**.
- **Temporal Relations**: Two events have temporal order, like **one event** happens after **another event**.

More examples can be found below.

A good way to check the event commonsense plausibility of a response is to think about the **events** in the response and check whether their **relations** are commonsensical. For example,

- Dialogue History:  
A: Is there any clue?

**Response:**

B: If I had a clue, I'd be happy to help.

In the **response**, there are two events, **"have a clue"** and **"be happy to help"**. From "If I ..., I would ...", we can know the **response** assumes there is **a relation of "Prerequisite"** (you may also think of it as a **temporal relation**). This relation is sensible according to our commonsense. **So, we think this response is absolutely commonsensical regarding the event commonsense.**

Since the **response** is highly related to its previous utterance, you also need to consider them together. Similarly, you can think about the **events** in the previous utterance and check their **relations** to the **events** you have thought of in the response.

We provide more examples below. You may look at them if you are still not clear about how to evaluate. (Click "More Examples (click to expand/collapse)" below to see them!)

Figure 7: Annotation guideline for event commonsense scoring.

Dataset	Size	Average score	Average # tokens in target response
DECO	300	3.39	17.6
ConTurE Subset	100	2.88	15.7

Table 9: Statistics of collected dialogue in event commonsense evaluation datasets.

Gurevych, 2019). We use BART as the backbone model to align with the COMET model used in our method. For hyperparameter search, we fine-tune the cross-encoder for 10 epochs with the batch size of 8 and the learning rate from {1e-5, 3e-5, 5e-5}.

- **Cross-encoder (COMET)**: We use the off-the-shelf COMET model trained on ATOMIC<sub>20</sub><sup>18</sup> as the backbone model. Other implementation details are the same with **Cross-encoder**.

- **MLP regressor**: Since the code of Zhou et al. (2021) is not publicly available, we produce the results using our own implementation based on scikit-learn<sup>19</sup>. Our implementation is available in our released codebase.

For event-relation extraction baselines, the im-

<sup>18</sup>[github.com/allenai/comet-atomic-2020](https://github.com/allenai/comet-atomic-2020)

<sup>19</sup>[https://scikit-learn.org/stable/modules/generated/sklearn.neural\\_network.MLPRegressor.html](https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html)



	xIntent	xNeed	xReact	oReact	xWant	oWant	xAttr	xEffect	oEffect	HinderedBy	IsAfter	HasSubEvent
Train (few-shot)	20	20	24	30	22	22	25	37	22	29	26	30
Test	37	35	21	50	29	20	46	71	32	20	61	45

Table 10: Statistics of the collected event-relation tuples. We collect around 30 tuples for each relation from DECO training set to train the event-relation extractor in the few-shot learning setting. The test set of DECO is thoroughly annotated to provide a test set for the joint event-relation extraction task.

plementation details are as follows:

- **ASER Extractor:** We use their provided code<sup>20</sup> to extract events. The neural classifier for relation prediction is trained on the human annotated tuples in DECO training set.
- **CSKB Search:** We search tuples related to the target response and its previous response in ATOMIC<sub>20</sub><sup>20</sup> following the CSKB search pipeline described in Zhou et al. (2021). A potential concept set is built from the utterances by identifying nouns, verbs, and adjectives that are not stopwords through part-of-speech (POS) tagging and lemmatizing them. We return tuples whose head and tail both contain words in the concept set as the search result.

For CSKB population baselines, we use the implementation in Fang et al. (2021a)<sup>21</sup>. For the backbone model, we use RoBERTa<sub>LARGE</sub> which has roughly the same parameter budget with COMET. We train all models for 1 epoch with the batch size of 64. The learning rate is searched from {1e-7, 1e-5, 1e-3} on the validation set.

## E.2 Applying ACCENT to CSKB Benchmark

In §5 Setup 3, we apply the compatibility test approach in ACCENT to a CSKB benchmark. Such an application is seamless because the compatibility test also assigns a score to each tuple, and tuples which receive higher compatibility scores are naturally more suitable to populate the CSKB. We train the COMET model on the positive samples in the training set of the benchmark dataset for 1 epoch with the batch size of 64. The learning rate is searched from {1e-7, 1e-5, 1e-3} on the validation set. Note that our approach does not require any negative sample in the training stage. It also does not need the event graph information provided in the benchmark dataset, but results in Table 4 shows

<sup>20</sup><https://github.com/HKUST-KnowComp/ASER>

<sup>21</sup><https://github.com/HKUST-KnowComp/CSKB-Population>

	Pearson	Spearman
COMET (neural)	0.14	0.25
ACCENT approach	<b>0.40</b>	<b>0.42</b>

Table 11: Correlations between human judgments and different compatibility test approaches with human-extracted tuples on DECO test set.

that our method outperforms baselines which require manually created negative data and take in graph information.

## E.3 Computational Resources

We run BERTSAGE and KG-BERTSAGE for the CSKB benchmark experiments on a single Nvidia V100 GPU with 32 GB memory where these models require large memory consumption in the run time. The rest of experiments is done on a single Nvidia A10 GPU with 24 GB memory.

Note that although we develop the ACCENT framework based on large language models, the only part which requires training is the T5 model (with 223M parameters) for event-relation extraction. As discussed in §3.2, the model is fine-tuned in a low resource setting and the training takes less than 0.5 GPU hour.

## F More Discussion of the Compatibility Test Approach in ACCENT

ACCENT checks whether an event-relation tuple  $(h, r, t)$  is compatible with the commonsense knowledge by comparing the similarity between  $t$  and commonsense tails generated by the Dynamic CSKB (COMET). Ablation results in Table 5 show that the compatibility test approach in ACCENT yields better performance than the “COMET (neural)” alternative which also uses the COMET model. To exclude the potential noise introduced by the automatically extracted tuples, we further compare these two methods using human-extracted tuples on DECO test set. Results in Table 11 demonstrate that the conclusion still holds under this experimental setting. Table 12 gives

Dialogue	Tuple	Human	COMET (neural)	ACCENT approach
A: I work in the bakery and eat all my favorite cupcakes. What do you do?	(PersonX makes a mistake, xEffect, PersonX gets fired)	/	0.33	0.63
<b>B: I actually just got fired for a mistake I made.</b>	(PersonX gets fired, isAfter, PersonX makes mistake)	/	0.12	0.68
	(PersonX gets fired, HasSubEvent, PersonX makes mistake)	/	0.18	0.66
	<b>Average</b>	0.80	0.21	0.66
A: Yeah winter is coming soon. It gonna be really cold.	(PersonX wants to live in a cold place, xIntent, PersonX intends to go full on winter)	/	0.06	0.59
<b>B: I know I know. I want to live in a cold place before I go full on winter.</b>	(PersonX goes full on winter, xNeed, PersonX lives in cold place)	/	0.53	0.39
	(PersonX goes full on winter, isAfter, PersonX lives in a cold place)	/	0.95	0.55
	(PersonX knows winter is coming, HasSubEvent, PersonX wants to live in a cold place)	/	1.63	0.69
	<b>Average</b>	0.40	0.79	0.56

Table 12: Examples of results given by different compatibility test approaches. Only  $u_{n-1}$  in the dialogue history is shown. Human annotations are normalized to the range of (0, 1) and scores given by “COMET (neural)” are scaled up by 100 times for better demonstration. Compatibility scores for each tuple are averaged to get the final score for the target response (in **bold** font).

two samples with a breakdown of tuple results. Compared with the compatibility scores given by ACCENT approach, the scores given by “COMET (neural)” are less comparable among tuples with different relations, thus making this method unsuitable for ACCENT framework.

## G Error Analysis

We conduct a qualitative analysis of the event-relation component in ACCENT. Table 13 shows some examples of the extracted tuples. While most of the head and tail events are nicely composed and capture the major information in the given text, they are not perfect. Multiple participants involved in the dialogue further increase the difficulty of the task. We note that the model sometimes confuses the multiple participants in the dialogue and makes mistakes when using “PersonX” or “PersonY”. For example, in the third sample of Table 13, the model confuses different participants since the subjects of “raise money” and “go tomorrow for treatment” should be different. Such confusion will lead to the wrong tuples which cannot truly reflect the meaning of the dialogue. Also, identifying relation from the given dialogue is challenging. Although we include negative samples (dialogues that do not have a certain relation) when fine-tuning T5, errors still exist (e.g., the tuple with “xAttr” relation in the third sample of Table 13).

---

**Automatically extracted event-relation tuples**

---

Dialogue History:

A: Cool. Are you religious? I have not told my parents I am wiccan yet.

B: I am the wife of a pastor. Spirituality is important in our home.

A: Oh. I grew up in a un religious home.

B: God is always around, I hope he does not mind that I do not like cleaning.

Target Response:

A: I'm not religious. I'm a Christian.

(PersonX not be religious, xAttr, PersonX be a Christian), (PersonX grew up in un religious home, xEffect, PersonX be a Christian), (PersonX like cleaning, isAfter, PersonX be a Christian), (PersonX like cleaning, HasSubEvent, PersonX be a Christian), (PersonX like cleaning, HinderedBy, PersonX be a Christian)

---

Dialogue History:

A: I'm dreadfully sorry for you.

B: Thank you very much for your kindness.

A: Is there any clue?

Target Response:

B: If I had a clue, I'd be happy to help.

(PersonX have a clue, xAttr, PersonX be happy to help), (PersonX has a clue, xReact, PersonX be happy to help), (PersonX has a clue, xEffect, PersonX be happy to help), (PersonX asks PersonY for a clue, oEffect, PersonY is happy to help), (PersonX help PersonY, HinderedBy, PersonX have a clue)

---

Dialogue History:

A: you mean if they catch the murderer. I'm scared stiff about going into the city center at night.

B: There must have been some good news in the newspaper. I can't believe that none of the news stories make you happy or excited.

A: Well, there was one good piece of news. You remember the local girl who was dying of a rare blood disease?

B: Yes. Her parents were raising money to have her treated in the united states.

Target Response:

A: Well, they've got the money and she's going tomorrow for treatment.

(PersonX raise money to have PersonY treated in United States, xAttr, PersonX go tomorrow for treatment), (PersonX raises money to have PersonY treated in United States, xReact, PersonX feels good), (PersonX raise money to have PersonY treated in United States, xEffect, PersonX go tomorrow for treatment), (PersonX raise money to have PersonY treated in United States, xWant, personX go tomorrow for treatment), (PersonX raise money to have PersonY treated in United States, xIntent, PersonX go tomorrow for treatment), (PersonX raise money to have PersonY treated in United States, oEffect, PersonY go tomorrow for treatment), (PersonX raises money to have PersonY treated in United States, oReact, PersonY feels happy), (PersonX raise money to have PersonY treated in United States, oWant, PersonY go tomorrow for treatment), (PersonX go to the United States, isAfter, The parents raise money to have the girl treated)

---

Dialogue History:

A: Yeah I know, he did a great job for the limited time, can you believe that 20th Century Fox let George Lucas keep the merchandising rights, that's what made him a billionaire, so crazy I don't think they would do that ever again.

B: I think I can believe it. Good for him, what does the corporation need with something they didn't invent anyways?

A: No I agree with you. Are you going to watch the Academy Awards by the way? They give out 24 Awards and I think that the first broadcast was back in 1930, which is wild

B: Yeah, so like the academy awards totally suck. Does anyone actually give a bleep what they think?

Target Response:

A: Well I can't speak for the other commenters but I'm sure there are people who give them a bleep can help winners go on to win awards and make money.

(PersonX give a bleep, HasSubEvent, PersonX help winners go on to win awards)

---

Table 13: Some examples of the automatically extracted event-relation tuples. These tuples function as the intermediate symbolic representation in ACCENT framework.

Instructions (click to expand/collapse)

Thanks for participating in this HIT!

In this task, you need to extract events *from the Response and its previous utterance* according to the given relation. The expected outcome consists of 3 parts: **Event A, Relation, Event B**

**Event A, Event B** Short phrases with a verb that may describe object properties, actions, etc.  
For example, in "I have just participated in the 3000-meter race, so I feel very tired now.", we can find two events, *I participate in the 3000-meter race* and *I feel tired*.

**Relation** How **Event A** relates to **Event B**.  
In the former example, there exists *a relation of "X wants" - "I feel tired" is the person's reaction to "I participate in the 3000-meter race"*.  
See detailed definitions and more examples below.

When composing **Event A / Event B**, please replace the person name or its pronoun with "PersonX". If there are additional people involved, replace them with "PersonY". For example,

<b>Source text</b>	<b>Event</b>
I have just participated in the 3000-meter race, so I feel very tired now.	PersonX participate in the 3000-meter race;
I'll invite my friends home.	PersonX feel tired
	PersonX invite PersonX's friends home

Sometimes, **Event A / Event B** may come from the sentence which doesn't have a subject. You need to add the subject yourself.

Here are the **Relations** you need to consider (more examples can be found below) :

<i>X effect</i>	as a result, X will
<i>X react</i>	as a result, X feels
<i>X intend</i>	because X intends

**Attention!** A couple NOTES:

- The two events should come from the **Response** or its **previous utterance**. If both two events are from the **Response**, please select "Single"; if one event is from the **Response** and another one is from its **previous utterance**, please select "Pair".
- Don't worry about whether the **Response** make sense or not. Simply find events and their relations according to the text.
- In a given sample, you may be able to find multiple (**Event 1, Relation, Event 2**) triplets. Please write down all of them.
- If you cannot find any tuple after considering all the given relations, feel free to report it (see below).

Relations (click to expand/collapse)


**Sample 1**

A: \${dlg\_history\_1}

B: **\${dlg\_response\_1}**

Click "Add new triplet" button to start working. Please write down all triplets you can find. If you really cannot find any triplet, please tick the check box below.

I have carefully consider **all relations** but I cannot find any triplet.

Relation	Event 1	Event 2	Evidence comes from
X effect X react X intend	<input type="text" value="Event 1"/>	<input type="text" value="Event 2"/>	Single Pair 

**Add new triplet**

Figure 8: Mechanical Turk template used to collect event-relation tuples. This template considers “xEffect”, “xReact”, “xIntent” relations.

## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
*Sections 6.2, 6.4 and Appendix H include the error analysis of the ACCENT framework. A separate section of "Limitations" is also included in Appendix A.*
- A2. Did you discuss any potential risks of your work?  
*We do not include potential risks of our work, since we believe that none of the components in our model and the dataset by itself can produce offensive results. The responses generated and augmented to DECO dataset are coming from previously proposed state-of-the-art models which are trained on datasets without profanity and inappropriate utterances. Other than that, our work is pertinent to evaluation and has less feasibility of potential risks.*
- A3. Do the abstract and introduction summarize the paper's main claims?  
*Abstract and Introduction summarize the paper's main claims. Three experimental setups and results show the ACCENT's superiority versus baselines.*
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

*Left blank.*

- B1. Did you cite the creators of artifacts you used?  
*In sections 5.1 and 6.3, we cite the CSKB population benchmark. Section 4.1 has the citations for each of the dialogue models used for the DECO response generation.*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
*Some dialogue sources we used come with licence. We discuss them in Appendix D.3.*
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
*Sections 5.1 and 6.3 use existing CSKB population benchmark that are compatible with the conditions that data was granted to be used. Our collected data is anonymized and section 4 describes the data use cases and statements.*
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
*Section 4 is mainly about data collection and creation, it discusses the data included in the collection process which are human judgments and it doesn't reveal the identity of users participated in the data collection process. The inputs for evaluation are coming from existing human-written conversational data and responses generated by dialogue models do not include inappropriate content.*
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
*Appendix D includes the language of collected dataset alongside the information about the data and its details.*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
*Section 4 and Section D.2 in the Appendix are about the size of datasets.*

**C  Did you run computational experiments?**

*Left blank.*

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
*Appendix E.3 reports the computational resources.*

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?  
*Yes we discussed them in Appendix E.*

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?  
*Sections 5 and 6 include the descriptive statistics about your results. Table 1 also demonstrates the results for ACCENT that all are significant ( $p < 0.05$ ).*

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?  
*Left blank.*

**D  Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?  
*Section 4 and Appendix D contain all the details about human annotations and conducted experiments.*

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?  
*The crowdsourcing platform is discussed in Section 4, and we discuss how we pay the annotators in Appendix D.1.*

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?  
*Yes, it was indicated in the AMT HIT pages.*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?  
*Not applicable. We did not include ethics review board for our data collection process.*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?  
*Appendix D.1 reports the basic demographic and geographic characteristics of the annotator population.*