What happens *before* and *after*: Multi-Event Commonsense in Event Coreference Resolution

Sahithya Ravi[1,2] **Chris Tanner**[3,4] **Raymond Ng**[1] **Vered Shwartz**[1,2]
[1] University of British Columbia
[2] Vector Institute for AI
[3] Massachusetts Institute of Technology
[4] Kensho Technologies
{sahiravi, rng, vshwartz}@cs.ubc.ca, cwt@mit.edu

Abstract

Event coreference models cluster event mentions pertaining to the same real-world event. Recent models rely on contextualized representations to recognize coreference among lexically or contextually similar mentions. However, models typically fail to leverage commonsense inferences, which is particularly limiting for resolving lexically-divergent mentions. We propose a model that extends event mentions with temporal commonsense inferences. Given a complex sentence with multiple events, e.g., "The man killed his wife and got arrested", with the target event "arrested", our model generates plausible events that happen before the target event – such as "the police arrived", and after it, such as "he was sentenced". We show that incorporating such inferences into an existing event coreference model improves its performance, and we analyze the coreferences in which such temporal knowledge is required.

1 Introduction

The goal of cross-document event coreference resolution is to determine if various event mentions (e.g. *shot*, *gunshot*), across one or more documents, refer to the same event. Existing systems represent each mention within its context using a language model (Cattan et al., 2021a; Allaway et al., 2021), and train a scorer to predict if two mentions corefer, based on their lexical and contextual similarity.

While many coreferring mention pairs in event coreference datasets such as ECB+ (Cybulska and Vossen, 2014) are lexically and contextually similar, or even share the same lemma (Wolfe et al., 2015), the difficulty arises for dissimilar coreferring mentions. For example, in Figure 1, *spent* and *hospitalized* are coreferring. These mentions are not lexically similar, and are not often used in similar contexts. In this paper, we improve the ability of existing cross-document event coreference systems to resolve such challenging coreferring mentions, by providing additional context in the form of commonsense knowledge. We focus on two temporal commonsense relations — *before* and *after* — pertaining to typical events that happen before and after the target event. For instance, in Figure 1, we may infer that before Dalton was *shot*, a shooter loaded their gun. Similarly, we may infer that Dalton was hurt prior to his *hospitalization* and got discharged afterward.

Our first contribution is the development of a commonsense reasoning engine that can reason about these two temporal relations. Existing commonsense models (Gabriel et al., 2021a; Hwang et al., 2021) may generate such inferences for simple sentences with a single event, such as "Bryant Dalton was shot", but they do not support complex sentences with multiple events of interest (e.g. *shot*, *hospitalized*). Further, they may conflate the inferences for different events. We develop a multi-event commonsense model that considers the entire context and is capable of generating separate inferences for each target event in complex sentences.

As an additional contribution, we incorporate the inferences into the pairwise mention scorer of a cross-document event coreference system (Cattan et al., 2021a). We produce *before* and *after* inferences for each event mention. We then embed the inferences, either by attending each mention to its own inferences (intra-span) or to the other mention's inferences (inter-span).

The results confirm that commonsense inferences are useful for event coreference. Each of our model variants improves upon the baseline performance, with the intra-span version performing the best. We further analyze the successful predictions and interpret how the commonsense inferences help resolve difficult mention pairs. In the future, we plan to extend our multi-event inference engine to additional commonsense knowledge types and apply it to other discourse tasks, such as summa-

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1708–1724
May 2-6, 2023

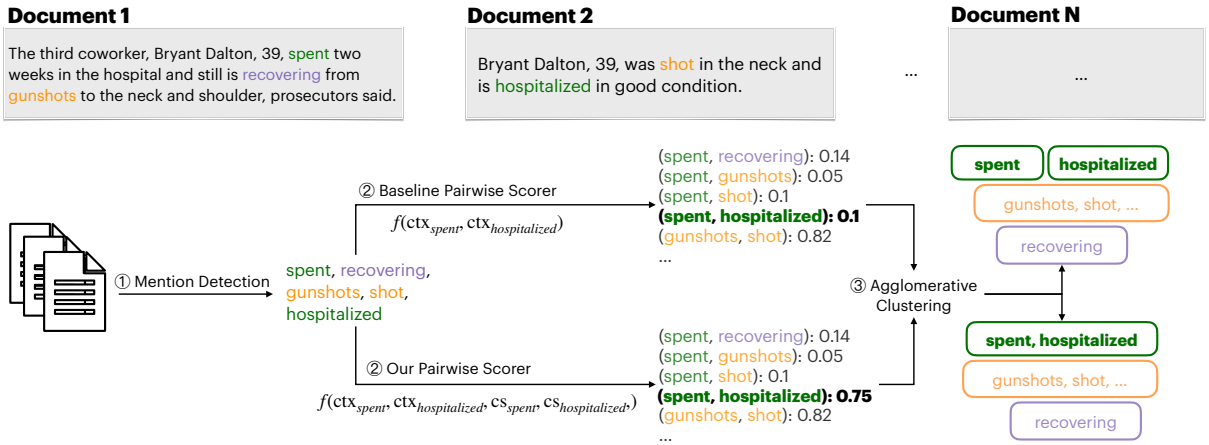

Figure 1: The architecture of our event coreference model, exemplified on a set of documents. The baseline pairwise scorer is based on contextual similarity (top), and we enhance it with temporal commonsense inference embeddings – indicated with *cs* (bottom). Such inferences help identify lexically divergent co-referring pairs, such as *spent* [time at a hospital] and *hospitalized*.

rization, dialogue, and story comprehension.[1]

2 Background

In this work, we improve the performance of a system for cross-document coreference resolution by incorporating temporal commonsense inferences pertaining to the events. We first provide background on event coreference resolution (§2.1). We then describe related work concerning event-centered commonsense (§2.2), along with approaches for using language models for data augmentation (§2.3).

2.1 Event Coreference Resolution

Event coreference resolution aims to cluster event mentions that refer to the same underlining real-world occurrence. Our work focuses on cross-document coreference resolution (CD), which aims to resolve mentions across an entire corpus of documents. In contrast, the problem of within-document coreference resolution (WD) only resolves mentions on a per-document basis. Event coreference is often performed jointly with *entity* coreference resolution, which concerns resolving mentions of people, locations, and organizations.

Datasets. In this paper, we use the ECB+ dataset proposed by Cybulska and Vossen (2014) and widely accepted as the standard benchmark for coreference resolution (CD). ECB+ contains 86 sub-topics, each of which concerns a specific news event. To introduce complexity and difficulty, each sub-topic is highly similar to – yet distinctly different from – exactly one other sub-topic. ECB+ includes both entity and event mentions; however, this work focuses solely on events. Table 1 shows event statistics from ECB+ corpus.

	Train	Dev	Test
Topics (subtopic-pairs)	25	8	10
Event mentions	3808	1245	1780
Event clusters	1527	409	805

Table 1: Statistics on the standard train/dev/test split of event coreferences in ECB+ dataset (Cybulska and Vossen, 2014)

Models. Recent approaches to CD event coreference often follow the architecture described in Figure 1. First, candidate event mentions are extracted from documents. Second, a pairwise scorer is trained to classify every pair of mentions as being coreferent or not. Finally, these scores are used to form distinct clusters of event mentions, typically using agglomerative clustering. Amongst these components, coreference models tend to mostly vary in their scoring approach (i.e., second component).

Early approaches relied on lexical and syntactic features (Yang et al., 2015; Choubey and Huang, 2017), or used semantic roles to encode the relationships between entities and events. Recently, Meged et al. (2020) improved performance by leveraging a resource of predicate paraphrases. Finally, Lai et al. (2021) incorporated entities, relations, and events extracted from a state-of-the-art information ex-

[1]The code is available here.

traction system. Generally, current state-of-the-art models often rely on pre-trained language models to compute a contextualized representation for each candidate mention, which serve as input to the pairwise scorer (e.g. Yu et al., 2020; Zeng et al., 2020; Cattan et al., 2021a; Allaway et al., 2021).

Our model is an enhancement of the model proposed by Cattan et al. (2021a). It targets both entity and event coreference resolution, and in an end-to-end fashion it performs mention extraction, pairwise scoring, and clustering (Figure 1). Mentions are represented by contextualized embeddings from RoBERTa (Liu et al., 2019). We chose to base our model on Cattan et al. (2021a) for two reasons. First, it is a simple model following the standard approach presented in Figure 1. Later approaches rely on hierarchical representations (Yadav et al., 2021a) or discourse coherence theory (Held et al., 2021). Second, it is based on RoBERTa and is more efficient and less memory consuming than the succeeding CDLM model (Caciularu et al., 2021) that is based on the much larger Longformer model (Beltagy et al., 2020).

More recently, Yadav et al. (2021a) built on Cattan et al. (2021a) by proposing a hierarchical approach to representing uncertainty of clustering event and entity mentions. The state-of-the-art models for cross document coreference are Caciularu et al. (2021), which models cross-text relationships by using larger context windows, and Held et al. (2021), which applies discourse coherence theory to coreference.

2.2 Event-Centric Commonsense

Commonsense reasoning helps humans bridge the gap between utterance and intended meaning. Reasoning about events has long been of interest to AI research. Schank and Abelson (1975) introduced "scripts" as a prototypical series of events, e.g. going to a restaurant is composed of ordering food, eating, and paying, and the participants: customer, waiter, and cook. Various methods have been proposed to learn such scripts from text (e.g. Chambers and Jurafsky, 2008; Pichotta and Mooney, 2014; Rudinger et al., 2015).

The ATOMIC knowledge base (Hwang et al., 2021; Sap et al., 2019) consists of 1.1M crowdsourced event-relation-event triplets pertaining to the causes, effects, and mental states of the event participants. To generate contextually-relevant ATOMIC-style inferences, Bosselut et al. (2019) developed COMET, a pre-trained language model fine-tuned on ATOMIC. COMET has shown promising results on tasks such as therapy chatbots (Kearns et al., 2020), persona-grounded dialogue (Majumder et al., 2020), figurative language interpretation and generation (Chakrabarty et al., 2020, 2022), and temporal ordering of sentences (Ghosal et al., 2021).

Several variants of COMET have been subsequently released. ParaCOMET (Gabriel et al., 2021a) adapts COMET to generate sentence-level inferences within the context of an entire paragraph. VisualCOMET (Park et al., 2020) generates ATOMIC-style inferences for images. Finally, the updated version of COMET (Hwang et al., 2021) extends the relation inventory and crowdsources more inferences. The additional inferences include the two temporal relations that are the most relevant to our work, "happens before" and "happens after".

2.3 LM-generated Data Augmentation

The success of using large pre-trained LMs in a few-shot setup for generation tasks has led to an increased interest in using such models to generate data for downstream tasks. Recent work augmented datasets by fine-tuning a pre-trained LM on real data, then generated new, silver-labelled instances (Anaby-Tavor et al., 2020; Papanikolaou and Pierleoni, 2020; Kumar et al., 2020). Similarly, the few-shot capabilities of GPT-3 (Brown et al., 2020) were leveraged to generate free-text explanations (Wiegreffe et al., 2022), semantically-related sentence pairs (Schick and Schütze, 2021), atomic event commonsense triples (West et al., 2022), and labels for various generation and understanding tasks (Wang et al., 2021). In this work, we fine-tune GPT-3 with minimal human supervision to generate additional contextual data pertaining to events.

3 Method

The architecture of our method is shown in Figure 1. We use the same clustering method as in Cattan et al. (2021a) but revise the pairwise scorer. Our goal is to improve the model's ability to resolve coreferences between mention pairs that are not lexically or contextually similar, but where one mention could be inferred from the other using commonsense knowledge and reasoning. Thus, we develop a commonsense inference engine (Sec 3.1)

Instructions: Read the context sentence and write at least **two** inferences for question 1 and **two** inferences for question 2. As shown in the examples, each inference is expected to be a short sentence between **5-10 words**.
Context: A publicist says Tara Reid has **checked herself** into rehab. **Question 1:** What typically happens before the event checked herself? **Question 2:** What typically happens after the event checked herself?

Figure 2: An example task on Amazon Mechanical Turk.

and use it to enhance the pairwise scorer (Sec 3.2).

3.1 Multi-Event Commonsense Inferences

We enhance the pairwise scorer with commonsense inferences regarding the events' temporal aspects. Specifically, we focus on plausible events that might have happened before or after the target event. For example, in Figure 1, after being hospitalized, the victim received treatment.

We found COMET and its variants to be ineffective for generating inferences for our task. COMET was trained on the ATOMIC knowledge base (Sap et al., 2019). As the name implies, events are atomic, i.e., comprise a single verb phrase. Conversely, the existing event coreference datasets are based on news articles, where sentences often contain multiple events. COMET predictions for document 1 (Figure 1) have no indication which verb they pertain to. Moreover, COMET predicts that what happens after document 1 is murder, which contradicts the fact that the victim survived and was taken to the hospital. ParaCOMET (Gabriel et al., 2021b) facilitates generating consistent inferences for multi-sentence paragraphs, but it was trained on the ROCStories dataset (Mostafazadeh et al., 2016), which is in the fiction domain and in which sentences are also simple.

To that end, we trained a new multi-event commonsense inference engine. Given a sentence with multiple events (such as document 1), and a target event (e.g. *hospitalized*), the goal is to generate what might have happened before and after the target event—in the context of the entire sentence.

Model. We base the inference engine on GPT-3 (Brown et al., 2020). While GPT-3 is not directly applicable to the task of event coreference (Yang et al., 2022), it has been shown to contain a wealth of factual and commonsense knowledge as a result of extensive pre-training. Our goal is to use this knowledge to generate event-centric commonsense inferences without requiring extensive training. GPT-3 is especially well-suited for this task, as it has shown remarkable performance in learning from fewer examples in a variety of tasks.

Data. As the first step in training a multi-event commonsense model, we crowdsourced annotations for 100 events – using the gold standard event mentions from the ECB+ training set. To include a wide range of topics, we selected the first four events from each of the 25 topics in the training set.

We presented workers with a sentence with one or more events, and asked them to describe what happens immediately before and after the target event. Figure 2 shows an example[2]. We obtained annotations from three workers for each sentence, and instructed workers to write at least two inferences for each relation. This yielded a total of 600 inferences ($100 \times 3 \times 2 = 600$). We carefully reviewed the data and removed a handful of inferences that were of poor quality (i.e., incomplete or irrelevant sentences, which amounted to roughly 5% of the annotations).

The annotation task was conducted on Amazon Mechanical Turk (AMT). To ensure the quality of annotations, we required workers to have previously completed 5,000 AMT tasks, and to have an acceptance rate of 98% or higher. We limited the worker location to the U.S. and Canada, and presented workers with a qualification test similar to the task. We paid 7 cents for each event.

Training. We fine-tuned GPT-3 on the collected inferences. The input and output format was as follows:

```
Context: <context>
Event: <event>
Before: <before>
After: <after>
```

Table 7 shows the format inputted into GPT-3 for training (top row) and inference (bottom row).

Inference To generate inferences, we prompt the fine-tuned GPT-3 model with the context and the event. We generate up to 150 tokens using top-p decoding (Holtzman et al., 2020) with a cumulative

[2]See Appendix E for the exact template.

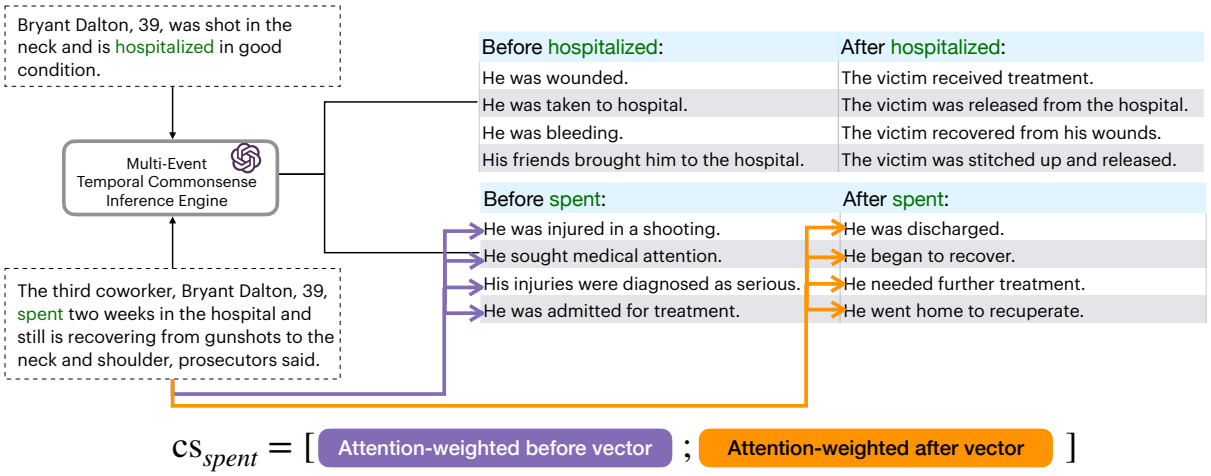

Figure 3: An illustration of the new additions to the pairwise scorer. We input each document into a GPT-3-based multi-event temporal commonsense inference engine, which outputs plausible events that happened before and after the target event (e.g. *spent*). For each temporal relation (i.e., *before* and *after*), we embed the corresponding inferences and compute an attention-weighted vector. We concatenate the *before* and *after* vectors to the mention representations as input to the pairwise scorer.

Context	Before	After
	Human-Written	
[...] Chris Weitz will direct the sequel to Twilight, New Moon, replacing Catherine Hardwicke.	They couldn't strike a [...] deal with Hardwicke.	Chris Weitz received an advance from the studio.
	The executive producer contacted Weitz's agent.	Chris signed the contract.
	Weitz's agent communicated the director's message.	His agent put out a press release.
	They decided to replace him.	Chris was happy.
[...] Chris Weitz will direct the sequel to Twilight, New Moon, replacing Catherine Hardwicke.	The director oversaw the hiring of shooting staff.	People watch the movie.
	The director oversaw several screen tests.	He gets paid.
	He writes a movie script.	The movie gets released in movie theatres.
	He needed to sign a contract.	The movie makes a huge collection.
	Model-Generated	
Lindsay Lohan checks into rehab at Betty Ford Center, rehires longtime lawyer Shawn Holley	She decided to change her life from bad to good.	She is treated for her addiction.
	She decided to seek help for her addiction.	She attended daily group therapy meetings.
	She is assessed by the staff at Betty Ford.	She ends up in the hospital.
	She is welcomed by staff members at Betty Ford.	She no longer has a problem.
Lindsay Lohan checks into rehab at Betty Ford Center, rehires longtime lawyer Shawn Holley	Lindsay needs advice on her case.	He gets a good pay.
	Lindsay needs legal counsel in her case.	He looks for another case.
	she fired her old lawyer.	she went through planning stages of her recovery.
	she got a new director.	she started to addiction treatment.

Table 2: Human-written (top) and model-generated (bottom) examples from our multi-event temporal commonsense inference engine. Some examples are slightly abbreviated for readability.

probability of $p = 0.9$. Table 2 provides examples of the training data (top part) and generated inferences (bottom part) of our multi-event commonsense inference engine. Note, both the human-written and model-generated inferences differ for different events belonging to the same context. For example, according to our model, *after* the event "Lindsay checks into rehab," a plausible inference is that "she gets treated for her addiction." Yet, *after* the event "she rehires her longtime lawyer," our model infers that "he gets a good pay." [3]

3.2 Inference-Enhanced Pairwise Scorer

Figure 3 shows the overall architecture of our commonsense-enhanced pairwise scorer. We follow Cattan et al.'s mention span representation for mention m_i:

$$\text{ctx}_i = [x_{START(i)}, x_{LAST(i)}, \hat{x}_i, l_i] \quad (1)$$

[3] We also experimented with prompting GPT-3 in a few-shot setup (Sec 5.3).

where x_j corresponds to the RoBERTa (Liu et al., 2019) embedding of the jth token in the span. Each mention is represented as the concatenation of: the first ($x_{START(i)}$) and last ($x_{LAST(i)}$) tokens; an attention-weighted sum of tokens $\hat{x}_i$; and a feature vector denoting the length l_i.

To incorporate the commonsense inference, we use the inference engine to generate up to $k = 5$ inferences for each of the before (b) and after (a) relations: $b_1...b_k$, and $a_1...a_k$. We describe below the representation of the relation vector using *after* as an example. The representation of the *before* relation is identical. We first compute the contextualized representation of each inference similarly to the span representations in Equation 1.

We then stack all the contextualized representations of the inferences:

$$\overrightarrow{A_i} = [\text{ctx}_{a1} \ ... \ \text{ctx}_{ak}] \tag{2}$$

and input them into a single head attention layer, which produces a single attention-weighted vector for the *after* relation.

In the context of the pairwise scorer, consider that we have two mention spans m_i and m_j and their corresponding *after* inference representations A_i and A_j. We implement two variants of the attention mechanism:

1. *Intra-span*, where the attention is between the mention span m_i and the corresponding inferences $\overrightarrow{A_i}$. This is exemplified in Figure 3, where attention is computed between the mention span of *spend* and the inferences corresponding to the same document. The query vector is the mention span ctx_i, and the key vector is the contextualized *after* vector $\overrightarrow{A_i}$. The idea behind this method is to emphasize inferences that are the most relevant to the given mention and provide additional context.

2. *Inter-span*, where the attention is between the mention span m_i and the inferences generated for the context of the other mention, $\overrightarrow{A_j}$. For example, in Fig 3, this would mean the purple and orange arrows originating in document 1 would need to be moved to the top row of inferences, corresponding to document 2. The query vector is the span ctx_i, and the key vector is the contextualized *after* vector $\overrightarrow{A_j}$. The goal of this method is to emphasize inferences that are relevant to the other mention, and to bring lexically divergent mentions closer.

In both cases, this leads to an attention-weighted commonsense vector for each of the before and after relations, which are then concatenated to create a single commonsense vector $\text{cs}_i = [\overrightarrow{B_i}, \overrightarrow{A_i}]$ as shown in Figure 3. The input to the pairwise scorer for mentions m_i and m_j is therefore:

$$g_{i,j} = [\text{ctx}_i, \text{ctx}_j, \text{cs}_i, \text{cs}_j] \tag{3}$$

The scores from the pairwise scorer are then used to cluster mentions using agglomerative clustering, identically to Cattan et al. (2021a). Agglomerative clustering merges the most similar cluster pairs until their pairwise similarity score falls below a predetermined threshold.

4 Experimental Setup

4.1 Implementation Details

The implementation of our model is based on Cattan et al. (2021a). We use their official codebase[4] and modify it to support the additional components. Since we use gold event mentions to generate inferences from the multi-event commonsense inference engine (Sec 3.1), during both training and inference, we train and evaluate the coreference pipeline on gold mentions. During testing, we evaluate both GPT3 and the coreference system on new gold mentions that are not seen during training. This is in contrast to Cattan et al. (2021a) which learned to extract candidate mention spans and train the coreference system. However, using gold mentions is common practice among many coreference systems where the focus is on improving the pairwise scorer (e.g. Barhom et al., 2019; Yadav et al., 2021a). For a fair comparison, we report the baseline performance by re-running Cattan et al. (2021a) using gold mentions similar to the baseline used in Yadav et al. (2021b). We compare this baseline to two variants of our model, based on intra-span and inter-span attention (Sec 3.2). We train all model versions using 15 different random seeds, and we report the average performance.

For our GPT-3 based inference engine, we fine-tuned the *Davinci* model which we accessed via the OpenAI API.[5]. The hyperparameters for all the models are detailed in Appendix B.

4.2 Evaluation Setup and Metrics

The primary metric we use is the standard CONLL-F_1 implemented by Moosavi and Strube (2016)[6], which is the average of three metrics: B^3 (Bagga

[4] https://github.com/ariecattan/coref
[5] https://beta.openai.com/
[6] https://github.com/ns-moosavi/coval

Model	MUC			B^3			CEAFe			CONLL	
	P	R	F1	P	R	F1	P	R	F1	F1	Δ
Baseline	73.49	84.13	78.45	48.49	67.72	56.52	43.49	55.65	48.83	61.30 ± 0.31	-
Inter-span	74.19	84.6	79.07	50.06	68.17	57.73	44.13	55.96	49.35	62.05 ± 0.35	(↑ 0.75)
Intra-span	75.02	84.72	79.58	51.01	68.00	58.29	44.31	57.70	50.13	62.67 ± 0.24	(↑ 1.37)

Table 3: Topic-level performance for event coreference on the ECB+ test set (with gold mentions, no singletons) - Baseline, Inter-span (multi-event commonsense), Intra-span (multi-event commonsense)

and Baldwin, 1998), MUC (Vilain et al., 1995), and $CEAF_e$ (Luo, 2005). We follow the evaluation setup used in recent work (Cattan et al., 2021a; Yadav et al., 2021a; Held et al., 2021; Cattan et al., 2021b) and evaluate all our models at the topic level. That is, each metric is computed for each topic separately and averaged across all topics. We also remove singleton clusters (clusters with a single mention) as they have shown to artificially boost the scores when using gold mentions (Cattan et al., 2021a).

5 Evaluation

We discuss the results on the event coreference task (Sec 5.1), the validity of the commonsense inferences generated by our inference engine (Sec 5.2), and present ablation tests (Sec 5.3).

5.1 Results

Table 3 shows the performance of the baseline and the inter-span and intra-span variants of the proposed multi-event commonsense models on event coreference on the ECB+ test set. Both of our proposed variants improve upon the baseline in terms of CONLL-F_1, with our intra-span model yielding an increase of 1.37 (± 0.24) points, and our inter-spam model yielding an increase of 0.75 (± 0.35). Overall, the improvement in performance indicates that the temporal commonsense inferences helped in resolving a considerable number of coreferences, which we analyze in more detail in Sec 6.1. In particular, both models improve upon the baseline precision across all metrics, with the intra-span model achieving the highest precision across all metrics. Error analysis of the best model (intra-span, Sec 6.2) shows that in some cases when mentions had similar (and possibly generic) inferences, the model falsely classified non-coreferring mentions as coreferring. We hypothesize that this error is more common for the inter-span model. When one mention's inference is lexically similar to the other mention, it would get more attention, increasing the likelihood of a false positive error.

5.2 Human Evaluation of Inferences

We manually evaluate the quality of the commonsense inferences generated by our inference engine (Sec 3.1). We randomly sampled 600 inferences from the validation set. We used the same AMT qualifications as in Sec 3.1 and paid 20 cents per HIT.[7] We presented three workers with a sentence and a target event, followed by the before and after inferences generated by the model. We asked them about the inference's (i) *likelihood*, i.e. how often would the given inference actually occur before (after) the target event; (ii) *relevance* with respect to the context; and (iii) *specificity* of the inference with respect to the target event. Table 4 presents the results. As expected, the generated inferences were almost always relevant to the corresponding event contexts. The majority of inferences (78.8%) were specific to the target event, but there was a significant percent of moderately specific inferences (19.4%) that often pertained to other events in the context. Finally, the majority of inferences either always (58%) or sometimes (36.1%) happen before or after the target event. These results reconfirm the extrinsic gains in Sec 5.1, and suggest that the inference engine may be useful for other NLP tasks. The inter-annotator agreement in terms of Fleiss kappa for the three metrics are as follows: Likelihood = 0.71, Relevance - 0.65, and Specificity - 0.84 (substantial agreement).

5.3 Ablation Tests

In Sec 3.1, we argued that COMET is insufficiently accurate for complex sentences with multiple events. To collect evidence, we replace our GPT-3 based commonsense inference engine with COMET and re-train the event coreference model. We used the newest COMET version (Hwang et al., 2021), along with beam search to decode the top 5 inferences for each relation type (before/after), ranked based on the model's confidence.

In addition, to justify fine-tuning GPT-3, we also

[7] See Appendix E for the HIT template.

	Metric	% High	% Moderate	% Low
1.	Likelihood	58.0	36.1	6.1
2.	Relevance	97.3	0.0	2.7
3.	Specificity	78.8	19.4	1.8

Table 4: Human evaluation results for the inferences generated by our commonsense inference engine.

Model	Inter-span	Intra-span
Baseline (no inf.)	61.3 ± 0.31	
COMET	61.51 ± 0.21	61.39 ± 0.32
GPT-3 few-shot	61.59 ± 0.26	61.64 ± 0.35
GPT-3 FT (ours)	62.05 ± 0.35	62.67 ± 0.24

Table 5: CONLL-F_1 performance on the ECB+ test set using different event commonsense knowledge sources.

replace our multi-event commonsense inference engine with a few-shot version of the model. We randomly sampled 8 of the human-written inferences (Sec 3.1) to prompt GPT-3, and we used the same instructions to prompt it to generate before and after inferences. In all experiments, the rest of the model is as described in Sec 3.2.

Table 5 presents the ablation results. The COMET-based model shows a marginal improvement from the baseline, yet performs worse than the multi-event inference engine. The few-shot GPT-3 model performs better, but we discovered that more training data could improve the specificity and accuracy of the inferences. Finally, our fine-tuned GPT-3 inference engine outperforms all models, thanks to its explicit training on multi-event inferences.

6 Analysis

6.1 Attention Scores

Figure 4 presents an example of a mention pair (*drunken driving*, *DUI*) that was incorrectly predicted as non-coreferring by the baseline and correctly predicted as coreferring by the intra-span model. The inferences for each mention are sorted and highlighted according to their corresponding attention weights. The highest scoring *before* inference for the first mention, "Jamal is drinking and driving", and the second inference for the second mention "Wiliams drank alcohol", are similar, which likely contributed to recognizing the coreference. Figure 5 similarly shows an example that was incorrectly predicted as non-coreferring by the baseline and correctly predicted as coreferring by the inter-span model. Here, we can clearly observe the interplay between the second mention *drove off* and the inferences of the first mention *hit* related to driving and fleeing from the scene. The lexical and contextual diversity of these mentions necessitates commonsense inferences to resolve the coreference. Appendix C provides a second set of examples.

Category	%
① Lack of Structure	29.5
② Generic Inferences	24.6
③ Insufficient Knowledge	19.5
④ Incorporation	18.1
⑤ Attention	8.3

Table 6: Error analysis of the intra-span model.

6.2 Error analysis

We analyze the errors in the best version of our model (intra-span). 95% of the errors made by this model overlapped with the errors made by the baseline, and only 5% were newly-introduced. We sampled 100 errors from the validation set and manually categorized them into the following categories, detailed below and quantified in Table 6. See Appendix D for examples from each category.

① **Lack of Structure**: Similar or identical mentions may refer to different events, as in "*Jackman* hosting the Academy awards" vs. "*Ellen* hosting the Oscars". Previous work incorporated semantic roles into the mention representation to identify such cases (Barhom et al., 2019). Our baseline model, as well as the inferences from our proposed approach, do not explicitly incorporate any linguistic structure, which results in these errors.

② **Generic Inferences**: The generated commonsense inferences are not specific enough with respect to the target event. This causes both false positive errors, when a pair of non-corefering mentions have similar generic inferences; and false negative errors, when coreferring mentions have dissimilar generic inferences.

③ **Insufficient Knowledge**: The inferences are relevant to the target event, but don't contain all the knowledge required to resolve these coreferences.

④ **Incorporation**: The inferences and attention scores were accurate, but the model did not use them effectively during incorporation.

⑤ **Attention**: The model either attended too much to unnecessary inferences (weights close to 1) or ignored crucial inferences (weights close to 0).

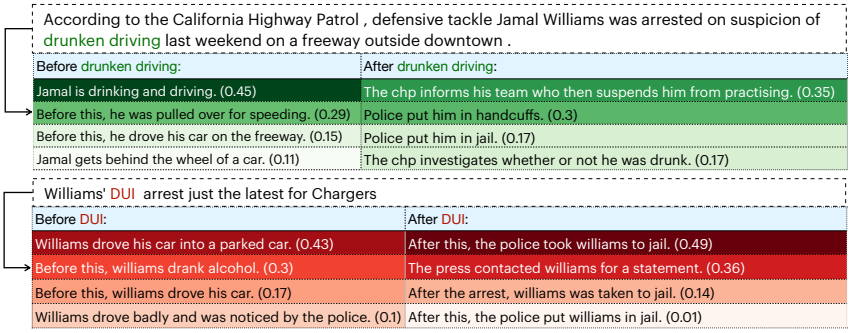

Figure 4: An example mention pair and the intra-span attention weights between the contexts and the inferences.

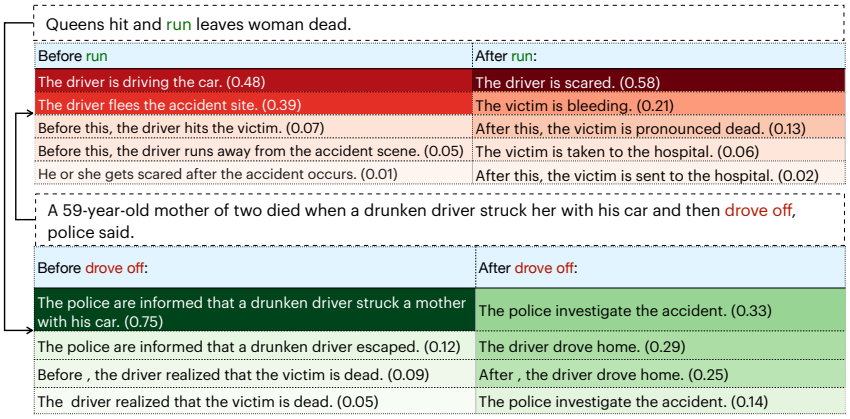

Figure 5: An example mention pair and the inter-span attention weights between the contexts and the inferences.

7 Conclusions

In this paper, we investigated the effect of injecting temporal commonsense knowledge in the task of event coreference resolution. By using event-specific inferences generated by our commonsense model, we improve the performance of a baseline model. Our analysis shows that the pairwise scorer attends to inferences that are beneficial in solving challenging coreferences. In the future, we plan to extend the multi-event commonsense model to additional relations, and to incorporate such knowledge into other discourse tasks.

8 Limitations

Data. As shown by Barhom et al. (2019), ECB+ suffers from annotation errors. In particular, the event coreference annotations are incomplete, which might lead to false positive errors for truly coreferring mention pairs. In this work, we intentionally addressed the edge cases in event coreference that haven't been addressed by prior research: lexically/contextually-divergent mentions. The number of such corefering clusters in ECB+ is small, and it has been shown that just clustering together mention pairs with the same lemma yields an F1 score of 42.3 on the dataset (Upadhyay et al., 2016). Further, our analysis of corefering pairs on the validation set revealed that only 11% of the pairs were contextually dissimilar (cosine similarity below 0.9), indicating that commonsense may impact only these cases. Unfortunately, this is the standard dataset for event coreference, but in the future, we could think of collecting a more challenging (and realistic) dataset.

Models. The accuracy of the commonsense model is primarily limited by the accuracy of inferences from GPT-3. Marcus and Davis (2020) tested GPT-3 on various types of commonsense reasoning and found mixed results for temporal commonsense. Our human evaluation in Sec 5.2 revealed that GPT-3 generates inferences that are not specific enough to the target event in 19.3% of the cases, which decreases performance as shown in Sec 6.2. We aim to address this in future work by building a more robust multi-event commonsense engine. Another error our model doesn't address concerns semantic roles, which happens when the main difference is in the person, time or location (e.g. two earthquake reports in different times and locations) (Barhom et al., 2019).

Evaluation. Since our commonsense engine was trained with gold event mentions, we used gold mentions to evaluate the coreference model as well. Using predicted mentions instead of gold mentions would provide a more realistic estimate of the performance of an event coreference system. With that said, our work focused on improving the coreference decisions; hence, we followed previous work and used the gold mentions (Barhom et al., 2019; Held et al., 2021).

9 Acknowledgements

This work was funded, in part, by the Vector Institute for AI, Canada CIFAR AI Chairs program, an NSERC discovery grant, and a research gift from AI2.

References

Emily Allaway, Shuai Wang, and Miguel Ballesteros. 2021. Sequential cross-document coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4659–4671, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, N. Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *AAAI*.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *LREC*.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021a. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107, Online. Association for Computational Linguistics.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021b. Realistic evaluation principles for cross-document coreference resolution. *ArXiv*, abs/2106.04192.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. Rˆ3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2017. Event coreference resolution by iteratively unfolding inter-dependencies among events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2133, Copenhagen, Denmark. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. 2021a. Paragraph-level commonsense transformers with recurrent memory. In *AAAI*.

Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021b. Discourse understanding and factual consistency in abstractive summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 435–447, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. STaCK: Sentence ordering with temporal commonsense knowledge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8676–8686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

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*.

William R Kearns, Neha Kaura, Myra Divina, Cuong Vo, Dong Si, Teresa Ward, and Weichao Yuwen. 2020. A wizard-of-oz interface and persona-based methodology for collecting health counseling dialog. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Tuan Lai, Heng Ji, Trung Bui, Quan Hung Tran, Franck Dernoncourt, and Walter Chang. 2021. A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3491–3499, Online. 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. *ArXiv*, abs/1907.11692.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online. Association for Computational Linguistics.

Gary Marcus and Ernest Davis. 2020. Experiments testing gpt-3's ability at commonsense reasoning: results.

Yehudit Meged, Avi Caciularu, Vered Shwartz, and Ido Dagan. 2020. Paraphrasing vs coreferring: Two sides of the same coin. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4897–4907, Online. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

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

Yannis Papanikolaou and Andrea Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *ArXiv*, abs/2004.13845.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*.

Karl Pichotta and Raymond Mooney. 2014. Statistical script learning with multi-argument events. In *Proceedings of the 14th Conference of the European*

Chapter of the Association for Computational Linguistics, pages 220–229, Gothenburg, Sweden. Association for Computational Linguistics.

Rachel Rudinger, Pushpendre Rastogi, Francis Ferraro, and Benjamin Van Durme. 2015. Script induction as language modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1681–1686, Lisbon, Portugal. Association for Computational Linguistics.

Maarten Sap, Ronan LeBras, 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 *AAAI*.

Roger C Schank and Robert P Abelson. 1975. Scripts, plans, and knowledge. In *IJCAI*, volume 75, pages 151–157.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *ArXiv*, abs/2104.07540.

Shyam Upadhyay, Nitish Gupta, Christos Christodoulopoulos, and Dan Roth. 2016. Revisiting the evaluation for cross document event coreference. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1949–1958, Osaka, Japan. The COLING 2016 Organizing Committee.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter West, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *NAACL*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Travis Wolfe, Mark Dredze, and Benjamin Van Durme. 2015. Predicate argument alignment using a global coherence model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.

Nishant Yadav, Nicholas Monath, Rico Angell, and Andrew McCallum. 2021a. Event and entity coreference using trees to encode uncertainty in joint decisions. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rohan Kumar Yadav, Lei Jiao, Ole-Christoffer Granmo, and Morten Goodwin. 2021b. Enhancing interpretable clauses semantically using pretrained word representation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 265–274, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent Bayesian model for event coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:517–528.

Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. 2022. What GPT knows about who is who. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 75–81, Dublin, Ireland. Association for Computational Linguistics.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2020. Paired representation learning for event and entity coreference. *ArXiv*, abs/2010.12808.

Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online). International Committee on Computational Linguistics.

10 Appendix

A Multi-Event Commonsense Inference Engine

B Hyper-Parameters

Table 8 shows the hyperparameters used by all our models. It took an average time of *120* minutes to run the entire pipeline for the inter-span and inter-span versions, and an average time of *80* minutes for the baseline version. We used a single NVIDIA GeForce GTX 1080 Ti GPU for each run. It took 5 minutes to fine-tune the GPT-3 Davinci model and costed 170 USD for training and generating all the inferences.

```
Context: Rumored to be the front runner earlier in the week , Entertainment Weekly has now confirmed
that Chris Weitz will direct the sequel to Twilight , New Moon , replacing Catherine Hardwicke.

Event: replacing

Before: They could not strike a favorable deal with Catherine Hardwicke. They decided to replace him.
Before this, the film's executive producer contacted Weitz's agent. Before this, Weitz's agent
communicated the director's message.

After: Chris Weitz received an advance from the studio. Chris signed the contract. After, his agent
put out a press release. Chris was happy. END
```

```
Context: Lindsay Lohan checks into rehab at Betty Ford Center , rehires longtime lawyer Shawn Holley

Event: rehires
```

Table 7: Examples of the input format of the multi-event commonsense inference engine. Top: a training example is fed into GPT-3 with the inputs (context and event) and the outputs (before and after inferences). Bottom: a test example is fed with only the inputs (context and event).

Parameter	Value
Batch Size	128
Learning Rate	0.0001
Dropout	0.3
Optimizer	Adam
Hidden layer	1024
Attention heads	1

Table 8: Hyperparameters used by all three model versions-Baseline, Inter-span and Intra-Span

C Attention Scores

In Figures 6 and 7, we provide examples for mention pairs incorrectly predicted by the baseline and correctly predicted by the intra-span and inter-span models, repsectively, similarly to Sec 6.2.

D Error Analysis

Table 9 shows one example of each error category described in Sec 6.2.

E Crowdsourcing Templates

Figures 8 and 9 show the HIT templates used for obtaining inference annotations and evaluating generated inferences, respectively.

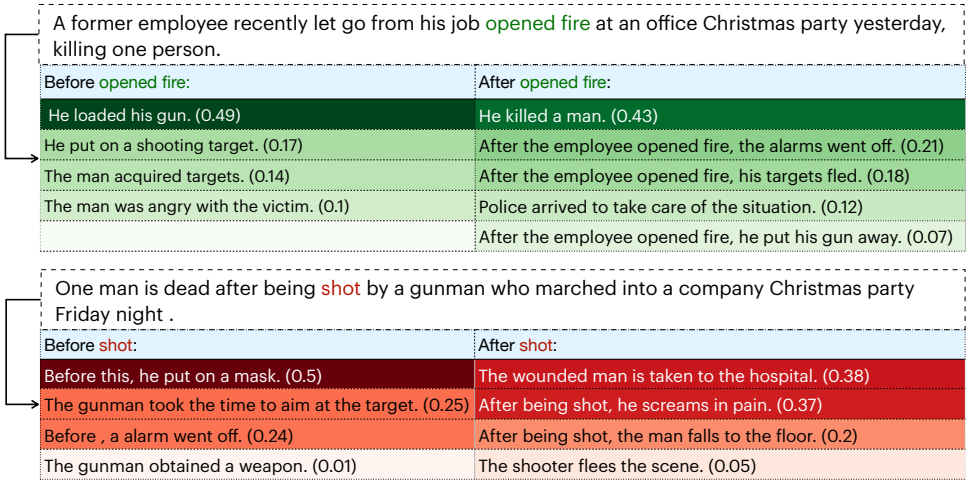

Figure 6: An example mention pair and the intra-span attention weights between the contexts and the inferences.

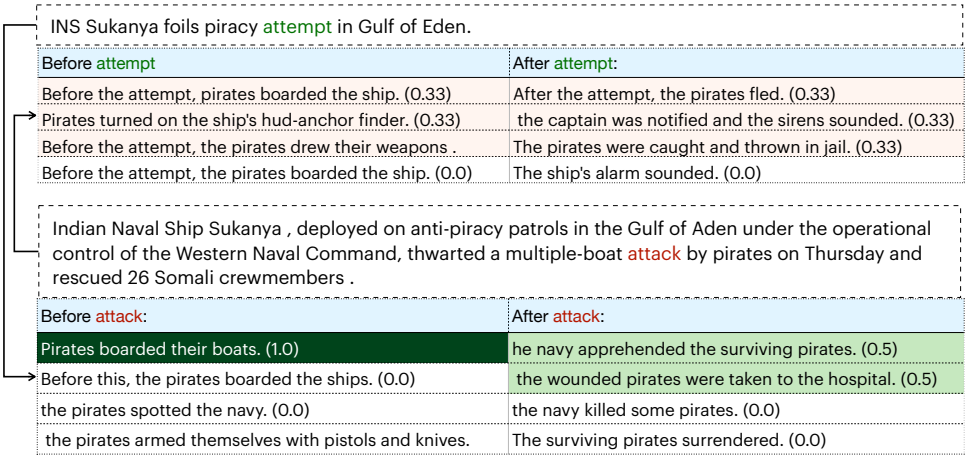

Figure 7: An example mention pair and the inter-span attention weights between the contexts and the inferences.

Context	Before	After
① **Lack of Structure: Different arguments (Robert Buckley vs Duncan Rait), similar event mentions and contexts.**		
Robert Buckley the second climber to die in the Aoraki - Mount Cook national park.	He slipped and fell down. (0.68) He set out to climb a mountain. (0.19) He was exhausted from the trek. (0.13)	The bodies of the climbers are found by the other climbers. (1.0) The families of the two climbers are notified. (0.0) the police investigate his cause of death. (0.0)
The day before Buckley's death another climber Duncan Rait, died after slipping and falling [...].	Before this, the climber slipped and fell. (0.5) The climber had an accident. (0.5) Before this, the climber sustained injuries. (0.0)	A rescue team went to look for the climber. (0.5) A funeral is held for the dead body. (0.5) the climber was pronounced dead. (0.0)
② **Generic error - Inferences of the first event (crash) are not specific and accurate.**		
Man charged with DWI , leaving scene after S . Rich Hill mother killed in crash : NYPD	They spotted a car on fire. (0.35) She swerved to avoid a cat crossing the road. (0.33) The driver got into an accident. (0.22) The car collided with the rich hill mother. (0.05) The rich hill mother is driving her car. (0.05)	They called the fire department. (0.41) Her family had to deal with the death. (0.34) Police arrived on the scene. (0.18) She was taken to the hospital. (0.07)
Woman Killed in Queens Hit - Run , Driver Charged	He changed his mind and decided to go forward with the plan. (0.46) A driver wants to kill the woman. (0.27) A driver sees the woman. (0.15) He gets scared and attempts to flee the scene. (0.06) He flees the scene. (0.06)	The woman is killed. (0.64) They file a case against the driver. (0.11) The driver gets worried about the consequence. (0.1) The woman is denied basic rights. (0.08) The woman is denied a burial. (0.08)
③ **Insufficient knowledge error- More knowledge may be beneficial (*e.g.*pre-requisites of events)**		
MSNBC is reporting that the Indian Navy claims they have captured 23 pirates in the Gulf of Aden	The navy ships noticed the pirates. (0.33) They boarded the ship. (0.33) The captain ordered an alert. (0.33) The navy ships surrounded the pirates. (0.0)	The captured pirates were taken to prison. (0.25) They will decide what to do with them. (0.25) the captain signaled the all-clear. (0.25) The navy notified the police about the capture. (0.25)
The Indian Navy on Saturday prevented pirates from attacking a merchant vessel[..] took 23 into custody.	They planned to attack the ship. (0.33) The pirates hid their weapons. (0.33) The navy received a distress call from the ship. (0.33)	The navy handed them over to the police. (0.33) The navy interrogate them (0.33) The navy took them to a different place (0.33)
④ **Incorporation error - Inferences seem relevant, but the model fails to use them.**		
5 Thoughts on Why the Academy Picked Ellen DeGeneres As Oscar Host	Ellen accepted to host the Oscars. (0.36) Ellen was practicing out ideas.(0.36) Ellen DeGeneres was selected as the host. (0.19) They academy contacted Ellen(0.09) The audience clapped for Ellen. (0.18)	Ellen feels happy(0.34) The host gets paid. (0.26) Ellen was given a plaque of honor. (0.22)
It will be her second stint in the job , after hosting the 2007 ceremony and earning an Emmy nomination for it	She practiced her speech. (0.32) She contacted her suppliers about a new gown for the show. (0.48) She was effective in her duties. (0.1) She was nominated for hosting the 2007 ceremony. (0.05)	The press contacted her for interviews. (0.55) She was very happy. (0.23) She informed her staff about the nomination. (0.12) She bought some new clothes. (0.1)
⑤ **Attention error: Increased attention on irrelevant inferences (first inference)**		
Woman Killed in Queens Hit - Run , Driver Charged	The driver came into contact with the woman. (1.0) The person driving a vehicle saw the woman and pursued, not caring about the person's safety. (0.0) The driver and the woman crossed paths.(0.0) The driver drove his vehicle at the woman. (0.0)	The driver flees the scene of the collision. (1.0) The woman is injured. (0.0) The woman is hospitalized. (0.0) The driver tried to hide his involvement in the crime. (0.0)
Cops : Queens Woman Killed In Hit - And - Run	A car flees the scene. (0.34) A car crashes into a dying woman. (0.26) They searched the area the car was spotted in. (0.21) They interviewed neighbors who might have seen them. (0.2)	They put out an alert to look for him. (0.28) They put out a press release calling for information. (0.28) They arrested him. (0.22) The criminal went to court (0.22)

Table 9: An example of each error category described in Sec 6.2

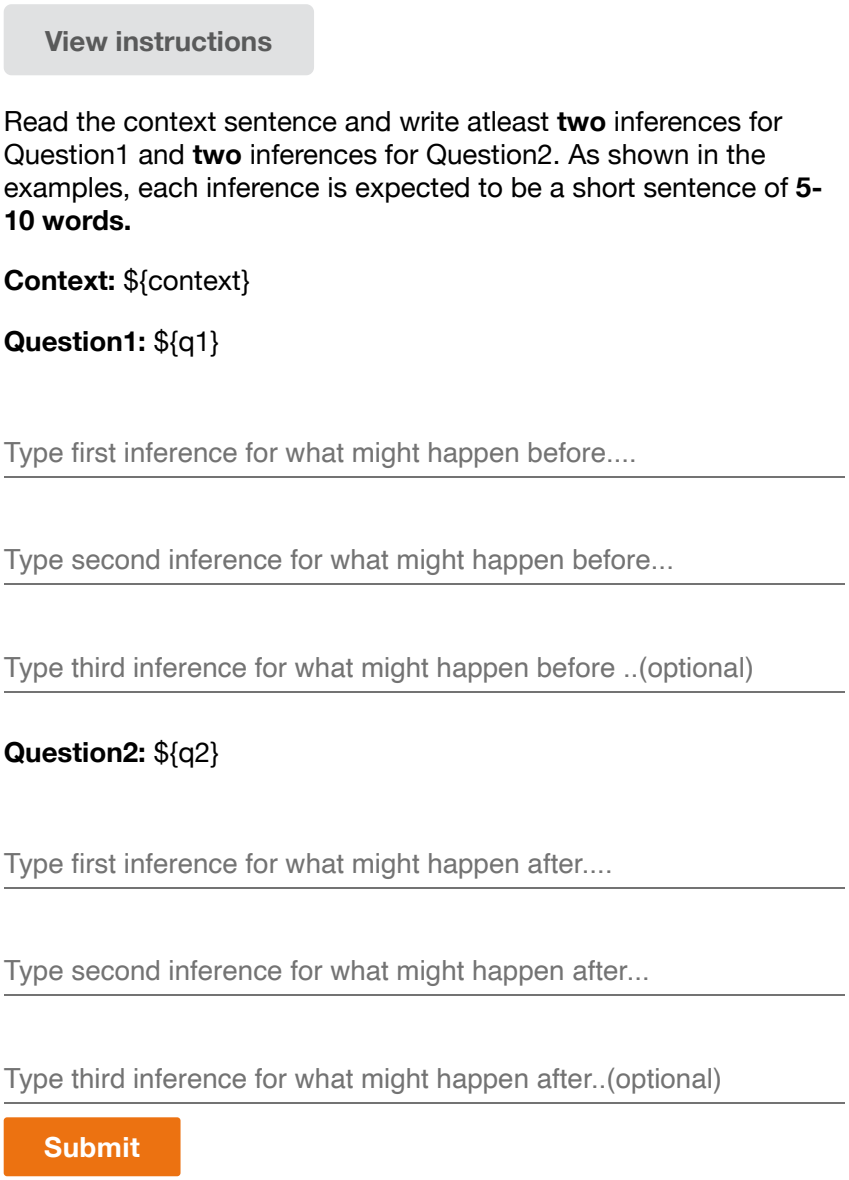

Figure 8: Crowdsourcing template for obtaining before and after inferences.

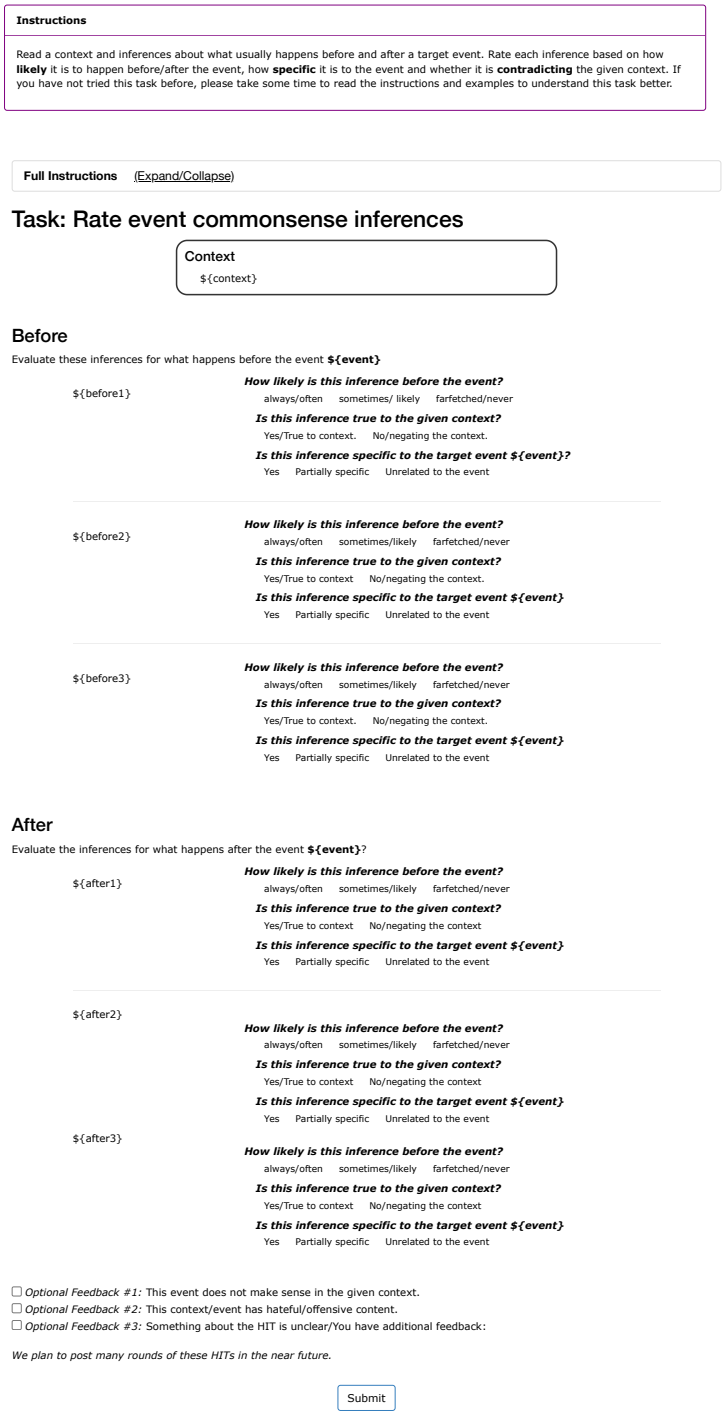

Figure 9: Crowdsourcing template for rating before and after inferences.