Paragraph-level Commonsense Transformers with Recurrent Memory

Saadia Gabriel[1,2], **Chandra Bhagavatula**[2], **Vered Shwartz**[1,2], **Ronan Le Bras**[2],
Maxwell Forbes[1,2], and **Yejin Choi**[1,2]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA
[2]Allen Institute for Artificial Intelligence, Seattle, USA
{skgabrie, mbforbes, yejin}@cs.washington.edu , {chandrab, vereds, ronanlb}@allenai.org

Abstract

Human understanding of narrative texts requires making commonsense inferences beyond what is stated explicitly in the text. A recent model, COMET, can generate such implicit commonsense inferences along several dimensions such as pre- and post-conditions, motivations, and mental states of the participants. However, COMET was trained on commonsense inferences of short phrases, and is therefore discourse-agnostic. When presented with each sentence of a multi-sentence narrative, it might generate inferences that are inconsistent with the rest of the narrative.

We present the task of discourse-aware commonsense inference. Given a sentence within a narrative, the goal is to generate commonsense inferences along predefined dimensions, while maintaining coherence with the rest of the narrative. Such large-scale paragraph-level annotation is hard to get and costly, so we use available sentence-level annotations to efficiently and automatically construct a distantly supervised corpus.

Using this corpus, we train PARA-COMET, a *discourse-aware* model that incorporates paragraph-level information to generate coherent commonsense inferences from narratives. PARA-COMET captures both *semantic* knowledge pertaining to prior world knowledge, and *episodic* knowledge involving how current events relate to prior and future events in a narrative. Our results show that PARA-COMET outperforms the sentence-level baselines, particularly in generating inferences that are both coherent and novel.

Introduction

Narrative understanding is a long-standing challenge in the field of natural language processing (NLP) (Charniak 1972; Winograd 1972). Arguably, the most crucial aspect of narrative understanding is the ability to make implicit commonsense inferences about entities and events in a story and refining them as the story unfolds (Pettijohn and Radvansky 2016; Williams, Lieberman, and Winston 2017; Rashkin et al. 2018; Qin et al. 2019). This ability in humans is seamless, yet essential for coherent understanding of narrative text. *Can NLP systems explicitly generate commonsense inferences, that a human might implicitly make while reading a narrative?*

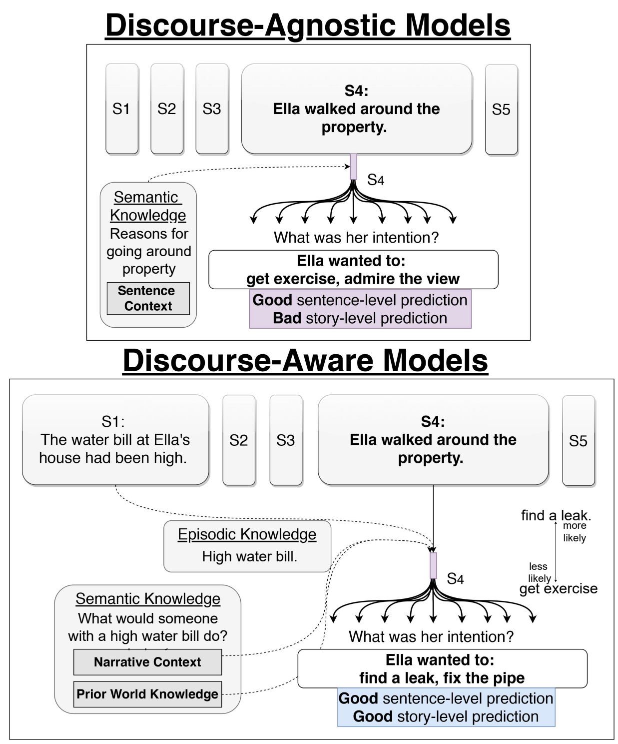

Figure 1: Discourse-agnostic models generate inferences relevant to the local context, but these generations can often be generic or incorrect at the narrative-level. Discourse-aware models take the rest of the context into account to make globally coherent inferences.

Being able to generate commonsense inferences has important practical implications. Commonsense Transformer (COMET, Bosselut et al. 2019), proposed recently, generates commonsense inferences for a given phrase or sentence, capturing pre- and post-conditions along nine inferential dimensions found in the ATOMIC (Sap et al. 2019) knowledge base.[1] The commonsense inferences generated by COMET have been effectively applied to downstream applications such as sarcastic comment generation (Chakrabarty et al.

[1]See Table 2 for a full list of inferential dimensions in ATOMIC.

2020), therapy chatbots (Kearns et al. 2020), abductive natural language generation (Bhagavatula et al. 2019), and automated story plot generation (Ammanabrolu et al. 2021).

However, the COMET inferences suffer from a major shortcoming – they are generated for a sentence in isolation and fail to account for the full *paragraph-level* narrative context. This often results in the generation of inferences that are inconsistent or unlikely when considering the previous narrative context. For example in Figure 1, given only the sentence *"Ella walked around the property,"* one might infer that she did this because she wanted to *"get exercise"* or *"admire the view"*. While such an inference is reasonable for the sentence in isolation, it is inconsistent given the full context – e.g., *"The water bill at Ella's house had been high. Ella walked around the property."* Instead, a more reasonable inference in light of the full context is that *"She wanted to fix a leak."*

We introduce the task of generating implicit discourse-aware commonsense inferences for narrative text, and present PARA-COMET, a transformer-based, controlled generation model for the task. Instead of collecting crowdsourced annotated data as direct supervision for this task, which is potentially expensive and challenging to scale, PARA-COMET is distantly supervised through sentence-level inferences obtained either from the COMET model or by heuristically matching a sentence to events found in the ATOMIC knowledge base. We define and use a coherence metric that measures the likelihood of each candidate inference in the context of the story to improve their paragraph-level consistency.

We show that PARA-COMET generates coherent discourse-aware inferences and performs better than discourse-agnostic baselines in both automated and manual evaluation. Yet, even the best model generates implausible inferences (23% of the inferences), and inferences that contradict the paragraph-level context (in 44% of the stories). This stresses the difficulty of the task and calls for further research. We release our models and data as an initial step towards advancing paragraph-level commonsense understanding.[2]

Background

Sentence-level commonsense inferences. A key component of our distant supervision approach is the availability of sentence-level commonsense inferences. The ATOMIC knowledge base (Sap et al. 2019) consists of such *if-then* knowledge about causes and effects, agents and themes of events, and their actions and mental states. An ATOMIC entry is encoded as a triplet $< e_1, d, e_2 >$, where e_1 is an event phrase, d is an inferential dimension and e_2 is the inference along the given dimension. ATOMIC defines nine inferential dimensions such as `xIntent`: the agent's intent, `oEffect`: the effect on the patient(s) etc. (See Table 2). The event e_1 and the inference e_2 are natural language templates consisting of variables `PersonX` for the agent and `PersonY` for the (possibly unknown) patient(s).[3]

While ATOMIC contains nearly 880K triplets, it is not nearly enough to capture the full range and generality of possible events, which is immeasurably vast and impossible to manually enumerate. Furthermore, due to lexical variability, events are rarely found as-is in ATOMIC. To that end, COMET (Bosselut et al. 2019) was developed as a transformer-based knowledge model trained on ATOMIC to generate commonsense inferences for a given phrase/sentence. Thus, both ATOMIC and COMET are natural candidates to obtain ***sentence-level*** commonsense inferences.

Reasoning about narratives. A related line of work to ours is *script learning*, that defines a structured representation for prototypical series of events (Schank and Abelson 1977). An event (e.g., going to a restaurant) is decomposed into components such as the participants (customer, waiter, cook, etc.), subevents (sitting down, asking for menus, etc.), and their various pre- and post-conditions. In later work, scripts were also referred to as "narrative event chains", and multiple methods to learn the narrative chains from raw text were developed (Chambers and Jurafsky 2008; Jans et al. 2012; Pichotta and Mooney 2014; Rudinger et al. 2015). Similarly, the *Choice of Plausible Alternatives* (COPA) task (Roemmele, Bejan, and Gordon 2011) proposes a benchmark for commonsense causal reasoning. It asks which of two alternatives has a causal relationship (either cause or effect) with a given premise. Finally, the temporal ordering of events is often studied along with typical times and duration (Kozareva and Hovy 2011; Granroth-Wilding and Clark 2016; Li, Ding, and Liu 2018; Zhou et al. 2019).

Types of commonsense inferences. While most commonsense work only pertains to non-situational semantic knowledge such as that captured by ConceptNet (Speer, Chin, and Havasi 2017), in this paper we focus on commonsense based on naive psychology, a core human ability that allows people to reason about mental states such as reactions, intents, goals and beliefs (Heider 1958) in particular situations. ATOMIC is specifically designed to capture such knowledge and we focus on such socially motivated commonsense, though our distant supervision approach and our proposed model are extensible to other knowledge bases and forms of commonsense.

Commonsense Inference with Discourse

Our work is motivated by the question: *can NLP systems explicitly generate commonsense inferences, that a human might implicitly make while reading a narrative?* To tackle this question, we formalize and introduce the discourse-aware commonsense inference task.[4]

[2]Code and data is available at https://github.com/skgabriel/paracomet.

[3]We refer to `PersonY` in ATOMIC as *patient*, one or more people who are affected or acted upon by the action of the verb. We don't make the semantic distinction between patient and theme.

[4]We use the term *discourse-aware* to refer to data/systems that use paragraph-level information. Similarly, *discourse-agnostic* systems only use sentence-level information.

Narrative	Dimension	w/o Discourse	w/ Discourse
Lenny was digging a hole in his yard to plant a tree. ... He jammed the shovel harder into the ground. **All of a sudden water started spurting out of the hole.**	`PersonX` needed to	be in a pool ✗	have a shovel ✓
Carla worked at the mall. **For her lunch break she ate at the food court.** ... Carla's co-worker bought her lunch.	`PersonX` needed to	be hungry ✓	drive to the foodcourt ✓
Sports day was always Emma's favourite day at school. ... A girl who moved to the school entered the 100m sprint. **Emma had never seen her ... thought she would be fine.**	`PersonX` wants to	practice more ✗	to win ✓
The water bill at Ella's house had been high. ... **Ella walked around the property.**	`PersonX` wanted to	admire the view ✗	find a leak ✓

Table 1: Examples generated from the models in this paper: a discourse-agnostic (sentence-level) baseline, vs. our discourse-aware PARA-COMET. We highlight the sentence that each inference was generated for in bold. Inferences are marked as plausible (✓) or implausible (✗).

Type	Dimension	Template
Causes	xIntent	PersonX wanted e_2
	xNeed	PersonX needed e_2
	xAttr	PersonX is seen as e_2
Effects	xWant	PersonX wants e_2
	xEffect	PersonX is likely e_2
	xReact	PersonX then feels e_2
	oWant	PersonY wants e_2
	oEffect	PersonY is likely e_2
	oReact	Others then feel e_2

Table 2: Natural language templates for ATOMIC dimensions.

Formally, given a narrative with T sentences $\{S_1, S_2...S_T\}$, the goal is to generate a set of commonsense inferences for the nine inferential dimensions (Table 2) for each sentence S_i. This set of inferences generated for S_i must also be consistent with the entire narrative. Maintaining consistency with the full narrative context requires reasoning about the relationship between past and future events.

Table 1 shows some examples of discourse-aware (paragraph-level) and discourse-agnostic (sentence-level) inferences. Sentence-level inferences are often inconsistent with the narrative. For example, the inference that a character needed "to be in a pool" when the earlier context shows they are gardening (first row in Table 1) or that a character wants to "practice more" when it has been established they are confident in their own abilities (third row).

Distant Supervision Approach

Sentence-level inferences (e.g. those obtained from COMET) are inadequate to train models for our proposed task and obtaining direct supervision of discourse-aware inferences may be prohibitively expensive or infeasible to collect in large quantities at an effective quality standard level. Therefore, we use distant supervision to loosely align sentences in a narrative to their discourse-aware commonsense inferences. First, we obtain discourse-agnostic inferences from either the COMET model or the ATOMIC knowledge base. Next, we filter out inferences that are inconsistent with the rest of the narrative (described in Section). Thus, we obtain *silver* standard training data for training models for our task. Additionally, we create a smaller-scale validation set by manually validating inferences through a crowdsourcing annotation task (Section).

Source of Narratives

The basis for our dataset are English stories from the ROCStories corpus (Mostafazadeh et al. 2016), which consists of 98K five-sentence stories authored by workers on Amazon Mechanical Turk. Understanding these stories requires commonsense and temporal inferences that we aim to capture. We split the original ROCStories train set into train, dev, and test sets in a 90/5/5 ratio.

Discourse-agnostic Inferences

We aim to generate the types of commonsense inferences defined by the ATOMIC knowledge base (Sap et al. 2019). We obtain discourse-agnostic inferences using either of the following approaches.

Heuristic: For each sentence S_i in the story, we get an initial set of candidate inferences R_i by extracting ATOMIC

Narrative	Inference	Relevant?
Natalie's favorite movie is The Wizard of Oz...	PersonX wanted: to see the film	✓
I was at the grocery store...**I see the lines were very long**...	PersonX then feels: relieved	✗
Jim wanted to learn Spanish. **He tried taking a class**...	PersonY/Others want: to catch up	✗
Our building had a summer bbq party today. The manager took photos...	PersonX wants: to enjoy the party	✓
Chris realizes that he rarely watches cable TV anymore. **He calls**...**to cancel**...	PersonX wanted: to be a good customer	✗
My grandparents lived in Alabama...**I miss traveling there**...	PersonX is seen as: sad	✓

Table 3: Examples from the distantly supervised dataset. We highlight the most relevant (i.e. potentially contradictory or supporting) sections in the story for each inference being considered.

tuples, $< e_1, d, e_2 >$, in which e_1 and S_i share either noun phrases or verb phrases. We repurpose the ROUGE metric (Lin 2004) to measure the surface-level relevance of a particular event e_1 to a sentence S_i. Specifically, we compute the ROUGE-1 F_1 score, which considers unigrams, and keep the top 10 inferences with respect to the score for each sentence and dimension.

Model-based: We use COMET to generate commonsense inferences for each sentence S_i in the story. We use beam search with a beam size of 10 to obtain a set of inferences for each sentence and dimension combination.

More details on the distant supervision data curation process are given in the Appendix.

From Discourse-agnostic to Discourse-aware Inferences

The inferences obtained by both heuristic and model-based methods (Section) only consider one sentence at a time. To improve coherence with the rest of the narrative, we filter the inferences that have a low *coherence* with the given narrative. Specifically, inspired by information theory (Shannon 1948; Hale 2001), we define coherence as a measure based on the cross entropy of the story tokens conditioned on a particular candidate knowledge inference. For a tuple $< e_1, d, e_2 > \in R_i$ matched to a sentence S_i, and a language model Θ, we compute the cross entropy loss of the tokens in the story, where $< d, e_2 >$ follow S_i: $CE(S_1, ...S_i, < d, e_2 >, ...S_5)$.[5] We use a transformer-based language model, and convert $< d, e_2 >$ to natural language using hand-crafted templates shown in Table 2.

In practice, we divide the dimensions into causes (`xNeed`, `xIntent`, `xAttr`) and effects (`xWant`, `xEffect`, `xReact`, `oWant`, `oEffect`, `oReact`). For cause inferences, we compute coherence with the previous and current sentences in the story. For effect inferences we use the full story. This allows us to effectively measure how well the extracted inferences may follow from past or predict future story events.

To ensure an equal distribution of inferences across dimensions, we order inferences by coherence score and keep the top 5 inferences for each sentence and dimension type. This filtering step is designed to reduce the number of contradicting inferences in our distant supervision corpus.

Validation Set

We validate a subset of the development set through crowdsourcing to obtain a gold evaluation set. We used Amazon Mechanical Turk and asked workers to judge the relevance of inferences for a given sentence within a story, leaving the interpretation of relevance to the best judgement of annotators.[6] Generally, we found that annotators adhered to a strict definition of relevance in which ambiguous inferences that may still be relevant to the story context at some point in the story timeline are labeled as irrelevant. See Table 3 for examples.

We randomly sampled 542 stories from the development set, and for each story we randomly selected a sentence and a dimension, and annotated the 5 inferences associated with them. We had 3 annotators judge each example, and used the majority vote to obtain a gold label. We filtered out low quality annotations by manually checking for workers with low inter-annotator agreement and frequently incorrect labeling.[7]

Our annotations yielded fair inter-annotator agreement of Fleiss' $\kappa = 0.338$ (Fleiss 1971) (p-value $< .001$). Despite the challenges of this task, this value is higher or comparable to prior work achieved for the evaluation of commonsense knowledge.[8] The final evaluation subset consists of 607 inferences, across all different dimensions, from 313 unique stories that were found to be relevant by multiple human annotators (34.29% of the inferences judged).

Model

We draw inspiration from the distinction between semantic and episodic memory (Tulving and Donaldson 1972), and consider implicit commonsense knowledge in two ways: 1) *semantic knowledge*, grounded in world knowledge and culture-specific social knowledge (e.g., "leaks lead to high water bills"), and 2) *episodic knowledge*, grounded in causal understanding and epistemic reasoning—i.e. reasoning that

[5] Here we define cross entropy loss as $CE(t_1, ..., t_n) = -\frac{1}{n}\sum_{i=1}^{n} log_2 p_\Theta(t_i|t_1, ..., t_{i-1})$.

[6] We restrict annotators to US only.

[7] These were done primarily by workers who spent less than 20 seconds on a HIT.

[8] $\kappa = 0.23$ in judging commonsense knowledge triplets in (Feldman, Davison, and Rush 2019) and between $\kappa = 0.289$ and $\kappa = 0.483$ in commonsense story generation in (Guan et al. 2020).

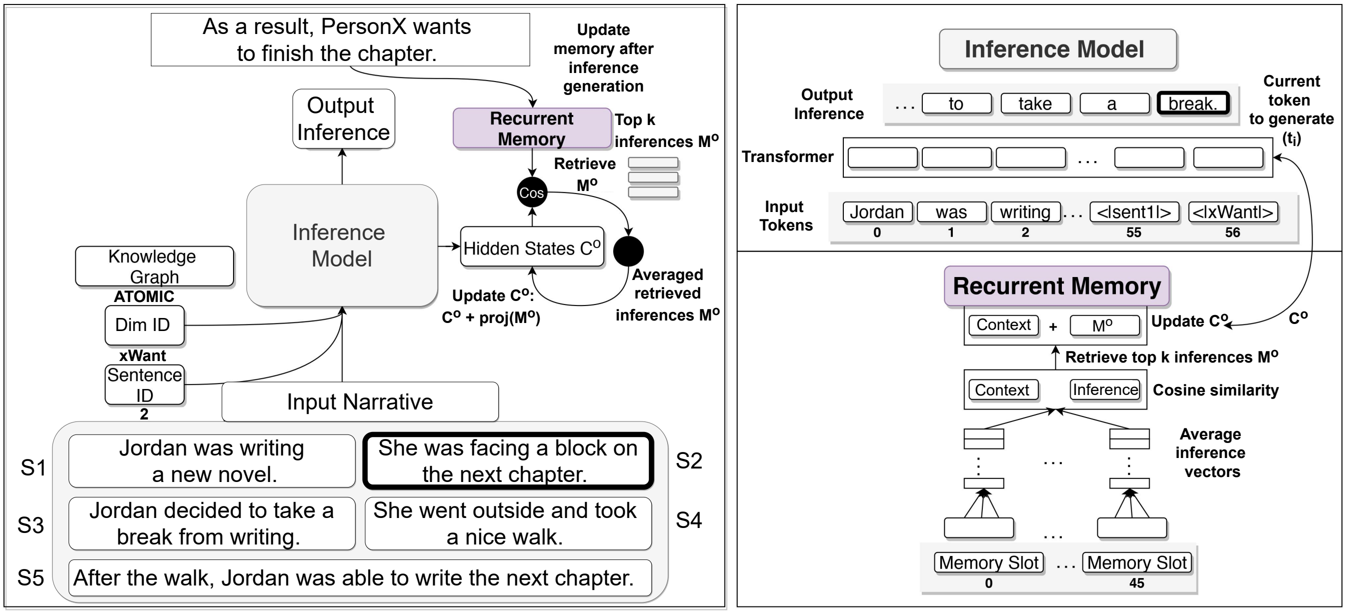

Figure 2: An illustration of PARA-COMET with a memory component. The model predicts an inference for a given sentence in the narrative (e.g., the second) and a requested ATOMIC dimension.

relates past events to current events (e.g., "if a person gets a high water bill, they will want to find out why"). We introduce two variants of the PARA-COMET controlled generation model: a memory-less model that focuses on semantic knowledge drawn from the context, and a model augmented with recurrent memory that allows us to explicitly incorporate episodic knowledge.

Figure 2 demonstrates generating inferences for a narrative using PARA-COMET with recurrent memory.

Memory-less model. Given a story context $c = \{S_1, S_2, \ldots, S_T\}$ of T sentences and a selected sentence S_i, we set the input to:

$$x = S_1 \mid\mid S_2 \ \ldots \ S_T \mid\mid s \mid\mid d \tag{1}$$

where s and d are special tokens. s represents the index of the selected sentence, while d represents the required dimension in ATOMIC. $||$ denotes concatenation. In the example in Figure 2, the input provided to the model is:

$$x = \text{Jordan was writing...} < |sent2| > \ < |xWant| >$$

We fine-tuned the base GPT and GPT2 transformer models (Radford et al. 2019; Radford 2018) to generate the expected output, which is an inference for the dimension d and sentence S_i.

Memory-augmented model. To incorporate inferences generated for other sentences in the story while generating inferences for a given sentence, we extend the model with a recurrent memory component, inspired by episodic memory. $M^m \in \mathbb{R}^{R^m \times L^r \times H}$ is the external memory, where R^m is either the maximum number of inferences per instance to store in memory (during training time) or the current number of instances (during decoding time), L^r is the maximum inference sequence length,[9] and H is the hidden state dimension.

The memory-augmented model takes as input a memory update matrix $M^u \in \mathbb{R}^{R^u \times L^r \times H}$, where R^u is the number of inferences used to update memory, and incorporates it into the memory matrix:

$$M^m = M^m \oplus f_{emb}(M^u) \tag{2}$$

$\oplus$ stands for matrix concatenation, and f_{emb} is an embedding layer trained jointly with the rest of the model. After the memory is updated, we average M^m across the token dimension to get $\theta^{mem} \in \mathbb{R}^{R^m \times H}$:

$$\theta^{mem} = \frac{1}{L^r} \cdot \sum_{l=1}^{L^r} M^{ml} \tag{3}$$

We denote the context representation obtained from GPT or GPT2's hidden state as $C^o \in \mathbb{R}^{L^c \times H}$, where L^c is the context sequence length. We average it across all tokens, obtaining $\theta^{ctx} \in \mathbb{R}^H$. We then prune the memory to the top-k most relevant inferences, measured by cosine similarity between the memory θ^{mem} and context vectors θ^{ctx}. The memory output $M^o \in \mathbb{R}^H$ is the average of the top-k inferences.

Finally, we reweigh the context representation C^o to consider the memory:

$$C^o = C^o + \text{proj}(M^o) \tag{4}$$

[9] We use a maximum memory size (R^m) of 45 inferences and a maximum sequence length of 100 tokens during training time. During decoding, we dynamically resize memory based on the number of inferences previously generated.

Where proj is a linear projection layer used to project the memory output into the same hidden dimensional space as the context representation.

At training time, the memory consists of previously extracted relations from our distant supervision, while at test time, it consists of previously generated inferences, recalling the model's prior decisions. For both PARA-COMET model variants, we minimize the cross entropy loss of the entire sequence (input and output).

Experimental Setup

Training Setup

All models are implemented using the Transformers package (Wolf et al. 2020), and trained for a maximum of 20 epochs. Training is performed using an Adam optimizer with linear warmup (Kingma and Ba 2015). We also simulate a batch size of 16 using gradient accumulation and an actual batch size of 4. The learning rate is $2*10^{-5}$ for GPT2. For GPT we use a learning rate of $6.25*10^{-5}$. All other hyperparameters follow (Radford et al. 2019; Radford 2018). We retrieve the top $k = 1$ inferences from memory.[10] We use the 124M parameter version of the GPT2 model.

Decoding Setup

For decoding, we use beam search with a beam size of $b \in \{1, 10\}$. The maximum decoding length is 50 tokens. Unlike at training time, where we take a single dimension for each sentence in each story, at decoding time we generate inferences from every dimension for every sentence. For both training and decoding, all experiments are run using 64 Intel(R) Xeon(R) Gold 6130 x86-64 CPUs at 2.10GHz and a Quadro RTX 8000 GPU.

Baselines

As a baseline, we use the COMET model, pre-trained on ATOMIC, to generate sentence-level inferences for each sentence in the story.[11] As an additional baseline, we use a retrieval model (BERT-KNN) based on the K-Nearest Neighbor search algorithm (k=1). We embed ATOMIC events using BERT (Devlin et al. 2019), then find the closest ATOMIC event node for each story sentence to get a set of matching inferences.

Evaluation

We report the performance of all models for automatic evaluation and the top 6 model variations (two COMET variations and four PARA-COMET variations) for human evaluation. For PARA-COMET, we report the variants with and without memory, trained on either the heuristic matching approach (PARA-H) or the model-based approach (PARA-M), as described in Section .

[10]For GPT2 we use memory during training and decoding. For GPT, we report results using training-only memory.

[11]See the original paper for details.

Human Evaluation

We follow a similar crowdsourcing setup to the validation presented in Section to measure the quality of generated inferences. We sampled 336 inferences from 56 unique stories. We show crowdworkers the full story, a specified dimension, and a generated inference. We specify the assignment of `PersonX` to the syntactic subject of the sentence.[12]

Following Zhang et al. (2017), we ask workers to judge the likelihood of inferences based on a 5-point Likert scale: obviously true (5), generally true (4), plausible (3), neutral or unclear (2), and doesn't make sense (1). Table 4 displays the percent of inferences judged as plausible or true (3-5), and plausible (3), and the average rating per inference (using majority voting).

Overall, PARA-COMET generations are scored with higher average ratings, between 3.05 and 3.44 points compared to 2.57 and 2.93 points for the COMET baseline variants. Specifically, the memory-augmented variants produced notably more plausible inferences than any other model. We observed that inferences in this category tend to be less obvious—e.g. restating information from the context, producing generic inferences—and recover plausible implicit inferences.

Automatic Evaluation

Similarity to the gold inferences. We follow the ATOMIC and COMET automatic evaluation setup using BLEU (Papineni et al. 2001), which measures the n-gram overlap between the generated and gold inferences.

Novelty. Following Jastrzebski et al. (2018), we compute novelty by measuring the percentage of generated inferences that do not appear verbatim in ATOMIC. We account for slight paraphrases by counting as novel the generated inferences that have an edit distance ratio of less than 0.95 with all ATOMIC events.

Discourse-level coherence. We use natural language inference (NLI; Dagan et al. 2013) as a proxy for measuring the narrative-level coherence of the predicted inferences. We define coherence as follows - at the very least, the story must not contradict any of the predictions, and it may possibly entail some of the predictions. We use the pretrained SemBERT model (Zhang et al. 2020), a variant of BERT augmented with explicit semantic role labels, to compute NLI labels (*entailment*, *neutral*, *contradiction*).

Table 5 provides a summary of the automatic evaluation results on the gold subset. The PARA-COMET variants outperform the sentence-level baselines across all metrics. The novelty results show that PARA-COMET models are capable of generating inferences that did not appear in the original ATOMIC knowledge graph. The memory-augmented models generated inferences that were

[12]We manually corrected incorrect parses such as those in which the subject of the sentence is not a person.

Model	Decoding	True or Plausible (3-5) (%)	Plausible (3) (%)	Avg. Rating
COMET	greedy	49.41	17.86	2.57
	beam-10	63.69	26.19	2.93
PARA-H	beam-10	68.45	22.62	3.21
PARA-H+mem	beam-10	66.67	27.98	3.05
PARA-M	beam-10	74.40	23.81	**3.44**
PARA-M+mem	beam-10	**77.38**	**31.55**	3.42

Table 4: Human evaluation results. We highlight the overall best performing model in bold. All PARA-COMET results are using GPT2 models.

Model	Decoding	BLEU-1	BLEU-2	Novelty	NLI
BERT-KNN	-	79.99	69.14	-	44.84
COMET	greedy	85.78	80.87	3.03	53.85
	beam-10	87.91	80.10	18.87	51.44
PARA-H (GPT)	beam-10	91.00	83.14	17.09	54.63
PARA-H+mem (GPT)	beam-10	90.99	83.43	16.09	56.23
PARA-M (GPT)	beam-10	91.03	83.06	12.56	52.72
PARA-M+mem (GPT)	beam-10	91.09	82.88	12.54	**59.42**
PARA-H (GPT_2)	beam-10	91.03	83.43	15.99	54.95
PARA-H+mem (GPT_2)	beam-10	**91.24**	**83.57**	14.39	56.23
PARA-M (GPT_2)	beam-10	89.44	81.89	**20.96**	54.63
PARA-M+mem (GPT_2)	beam-10	89.68	82.18	20.06	54.95

Table 5: Performance according to the automatic evaluation metrics. The best performing model for a specific PARA-COMET variant (GPT or GPT2) is underlined. We highlight the overall best performing model in bold. The NLI score is the percent of stories for which the model predicted entail or neutral.

more coherent with the story, reducing the percents of contradicting inferences from 46.15% (in COMET) to 40.58%. We find the GPT models generally have comparable or better performance to GPT2 models on automatic metrics, which we hypothesize is due to the fact GPT was pretrained on story text and has specific knowledge pertaining to implicit knowledge underlying narratives. Overall, we find that incorporating narrative coherence through either episodic knowledge from the recurrent memory mechanism and/or context from other story events improves BLEU-1 by up to 3.33 points and BLEU-2 by up to 2.70 points.

Case Study: Personal Narratives

To test the ability of the model to generalize to more complex narratives requiring further pragmatic reasoning (Sap et al. 2020), we sampled and manually evaluated a set of 111 story/sentence/dimension triplets from personal blog posts in the COSMOSQA machine reading comprehension test set (Huang et al. 2019). While these narratives tend to be of a similar or shorter length than ROCStories, they require more real-world understanding. They also contain nuanced descriptions of social interactions.

We found that our model is effective at predicting inferences in an unsupervised setting with 49.55 % of relations labeled as true and 20.72% of relations labeled as plausible (vs. 20.72% and 27.03% for COMET). We noticed that our model more frequently overcomes two major plausibility errors in unsupervised commonsense inference - *off-topic* and *temporarily inappropriate* predictions (see Table 6).

Story: Almost a month ago now, the radio station got struck by lightning. It fried the router and the cable modem. **[Before ϟ, PersonX wanted]** We got the new equipment right away.
COMET (beam-10): to be a good cook, to eat
PARA-M: to have better internet, to not be bothered
Story: I posted a moment ago regarding a girl I asked out... she said she would like to do something, but work made it difficult. That was a couple of weeks back... **[Next, PersonX will]**
COMET (beam-10): get married, get a divorce
PARA-M: be asked out, get rejected

Table 6: Examples of personal blog posts with commonsense model predictions. Here we assume PersonX to be the narrator of the blog post.

For example, our model is able to correctly predict the likely intentions of someone owning a router and cable modem (example 1), while COMET predictions incorrectly focus on meal preparation. COMET also sometimes makes relevant but farfetched predictions while our model's inferences are better situated within a narrative timeline (example 2).

Conclusion

We introduced a new task of discourse-aware commonsense inference over narratives. To target this task, we proposed a new model, PARA-COMET, trained using distant supervision, that captures narrative discourse.

Despite the challenges of the task, we demonstrated the effectiveness of our approach using both automatic and human evaluations. In particular, our models were able to generate more implicit and novel discourse-aware inferences. In the future, we are interested in exploring further extensions of our work to downstream paragraph- and narrative-level tasks that may benefit from access to commonsense knowledge.

Acknowledgments

We thank the anonymous reviewers for helpful feedback, as well as Maarten Sap, Hannah Rashkin, Eunsol Choi and members of the UW and AI2 communities for helpful discussions. This research was supported in part by DARPA under the CwC program through the ARO (W911NF-15-1-0543) and DARPA under the MCS program through NIWC Pacific (N66001-19-2-4031).

Ethics Statement

We note that the knowledge represented by current resources captures a mix of *general* commonsense which the majority of readers would find likely regardless of background (including factual commonsense) and *culturally-specific* commonsense which is only likely to some readers (i.e. is not the most logical conclusion for all readers). One likely contributor to this specificity of commonsense is the dependency on online crowdsourcing for annotation and generation of commonsense knowledge. A 2016 report from Pew Research[13] found that a sample of MTurk crowd-sourcing workers was heavily skewed demographically. This has an unintended side effect of enforcing a potentially harmful assumption that "commonsense knowledge" is only knowledge agreed upon by a specific demographic or cultural majority. Proposed steps for future work on discourse-aware commonsense inference include:

- Multicultural and multilingual commonsense datasets that capture a more distributional view of commonsense, allowing for both sociocultural overlap and disagreement about likelihood of relevant inferences.
- New evaluation metrics and frameworks for commonsense that consider likelihood rather than hard labels of relevancy. We begin to explore this with our categorical labeling of extracted commonsense knowledge.
- Social commonsense inference models that consider multiple audience points-of-view.

References

Ammanabrolu, P.; Cheung, W.; Broniec, W.; and Riedl, M. O. 2021. Automated Storytelling via Causal, Commonsense Plot Ordering. In *AAAI*.

Bhagavatula, C.; Le Bras, R.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; Yih, W.-t.; and Choi, Y. 2019. Abductive Commonsense Reasoning. In *ICLR*.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *ACL*.

Chakrabarty, T.; Ghosh, D.; Muresan, S.; and Peng, N. 2020. Rˆ3: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge. In *ACL*, pg. 7976–7986.

Chambers, N.; and Jurafsky, D. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pg. 789–797.

Charniak, E. 1972. *Toward a model of children's story comprehension*. Ph.D. thesis, Massachusetts Institute of Technology.

Dagan, I.; Roth, D.; Zanzotto, F.; and Sammons, M. 2013. *Recognizing textual entailment: Models and applications*. Morgan & Claypool Publishers.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Feldman, J.; Davison, J.; and Rush, A. M. 2019. Commonsense Knowledge Mining from Pretrained Models. In *EMNLP/IJCNLP*.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5): 378.

Granroth-Wilding, M.; and Clark, S. 2016. What happens next? event prediction using a compositional neural network model. In *AAAI*.

Guan, J.; Huang, F.; Zhao, Z.; Zhu, X.; and Huang, M. 2020. A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation. In *TACL*, volume 8, 93–108.

Hale, J. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *NAACL*.

Heider, F. 1958. *The Psychology of Interpersonal Relations*. John Wiley & Sons Inc.

Huang, L.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *EMNLP/IJCNLP*.

Jans, B.; Bethard, S.; Vulić, I.; and Moens, M. F. 2012. Skip n-grams and ranking functions for predicting script events. In *EACL*, 336–344. Association for Computational Linguistics.

Jastrzebski, S.; Bahdanau, D.; Hosseini, S.; Noukhovitch, M.; Bengio, Y.; and Cheung, J. C. K. 2018. Commonsense mining as knowledge base completion? A study on the impact of novelty. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*. Association for Computational Linguistics.

Kearns, W. R.; Kaura, N.; Divina, M.; Vo, C. V.; Si, D.; Ward, T. M.; and Yuwen, W. 2020. A Wizard-of-Oz Interface and Persona-based Methodology for Collecting Health

[13]https://www.pewresearch.org/internet/2016/07/11/turkers-in-this-canvassing-young-well-educated-and-frequent-users/

Counseling Dialog. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* .

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kozareva, Z.; and Hovy, E. 2011. Learning Temporal Information for States and Events. In *2011 IEEE Fifth International Conference on Semantic Computing*, 424–429.

Li, Z.; Ding, X.; and Liu, T. 2018. Constructing Narrative Event Evolutionary Graph for Script Event Prediction. In *IJCAI*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 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*, 839–849. San Diego, California: Association for Computational Linguistics. doi:10.18653/v1/N16-1098. URL https://www.aclweb.org/anthology/N16-1098.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

Pettijohn, K.; and Radvansky, G. 2016. Narrative event boundaries, reading times, and expectation. In *Mem Cogn 44, 1064–1075*.

Pichotta, K.; and Mooney, R. 2014. Statistical Script Learning with Multi-Argument Events. In *EACL*, 220–229.

Qin, L.; Bosselut, A.; Holtzman, A.; Bhagavatula, C.; Clark, E.; and Choi, Y. 2019. Counterfactual Story Reasoning and Generation. In *EMNLP*.

Radford, A. 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI technical report* .

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI technical report* .

Rashkin, H.; Bosselut, A.; Sap, M.; Knight, K.; and Choi, Y. 2018. Modeling Naive Psychology of Characters in Simple Commonsense Stories. In *ACL*.

Roemmele, M.; Bejan, C. A.; and Gordon, A. S. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Rudinger, R.; Rastogi, P.; Ferraro, F.; and Van Durme, B. 2015. Script induction as language modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1681–1686.

Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N. A.; and Choi, Y. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5477–5490. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.486. URL https://www.aclweb.org/anthology/2020.acl-main.486.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *AAAI*.

Schank, R. C.; and Abelson, R. P. 1977. *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Published by Psychology Press.

Shannon, C. E. 1948. The Mathematical Theory of Communication. *The Bell System Technical Journal* 27(3).

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Tulving, E.; and Donaldson, W. 1972. Episodic and semantic memory. *Organization of memory* .

Williams, B.; Lieberman, H.; and Winston, P. H. 2017. Understanding Stories with Large-Scale Common Sense. In *COMMONSENSE*.

Winograd, T. 1972. Understanding natural language. *Cognitive psychology* 3(1): 1–191.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Association for Computational Linguistics.

Zhang, S.; Rudinger, R.; Duh, K.; and Durme, B. V. 2017. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics* 5: 379–395.

Zhang, Z.; Wu, Y.-W.; Hai, Z.; Li, Z.; Zhang, S.; Zhou, X.; and Zhou, X. 2020. Semantics-aware BERT for Language Understanding. *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)* .

Zhou, B.; Khashabi, D.; Ning, Q.; and Roth, D. 2019. Going on a vacation takes longer than Going for a walk: A Study of Temporal Commonsense Understanding. 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)*, 3354–3360.