Automated Storytelling via Causal, Commonsense Plot Ordering

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl
Georgia Institute of Technology
{raj.ammanabrolu,wcheung8,d.tu,wbroniec3,riedl}@gatech.edu

Abstract

Automated story plot generation is the task of generating a coherent sequence of plot events. Causal relations between plot events are believed to increase the perception of story and plot coherence. In this work, we introduce the concept of *soft causal relations* as causal relations inferred from commonsense reasoning. We demonstrate C2PO, an approach to narrative generation that operationalizes this concept through **C**ausal, **C**ommonsense **P**lot **O**rdering. Using human-participant protocols, we evaluate our system against baseline systems with different commonsense reasoning approaches and inductive biases to determine the role of soft causal relations in perceived story quality. Through these studies we also probe the interplay of how changes in commonsense norms across storytelling genres affect perceptions of story quality.[1]

Introduction

Automated story generation is a standing grand challenge of AI. One of the central challenges of automated story generation is causal progression such that the events of the story follow from events that have come before. Many prior approaches to plot generation relied on symbolic planning (Lebowitz 1987; Gervás et al. 2005; Porteous and Cavazza 2009; Riedl and Young 2010; Ware and Young 2011)—reasoning directly about causal enablement in the form of predicate precondition and post-condition matching. While these systems can guarantee causal entailment between story events, these approaches also require extensive domain knowledge engineering and limited vocabularies of events and characters.

Machine learning approaches to automated story generation can learn storytelling and domain knowledge from a corpus of existing stories or plot summaries. This theoretically allows them to overcome the knowledge engineering bottlenecks. However, neural language model based approaches to automated story generation learn probabilistic relationships between words, sentences, and events and thus have difficulty modeling causal entailment between actions and events. Additionally, stories need to remain consistent with respect to genre and commonsense norms.

[1]Code found at https://github.com/rajammanabrolu/C2PO.

In this paper, we consider the challenge of automatically generating narratives that have recognizable causal entailment between events. Specifically, we approach the problem of story generation as a *plot-infilling* (Ippolito et al. 2019; Donahue, Lee, and Liang 2020) where an outline of plot points is extracted from a source then elaborated upon. We introduce the concept of *soft causal relations*, where causal entailment between story events does not need to be strictly logically consistent, but draws upon people's everyday commonsense understanding of whether one event tends to be preceded or succeeded by another.

We demonstrate an approach to story generation using soft causal relations in the C2PO (Commonsense, Causal Plot Ordering) system, which generates narratives via plot infilling using soft causal relations. Inspired by work on plot graph learning (Li et al. 2013), C2PO attempts to create a branching space of possible story continuations that bridge between plot points that are automatically extracted from existing natural language plot summaries. To create this branching story space, we iteratively extract commonsense causal inferences from the COMET (Bosselut et al. 2019) model of commonsense reasoning. Finally, once the space—a plot graph—has been constructed, we search the space for complete sequences.

Using human participation studies, we evaluate C2PO against baseline text infilling systems with different uses of commonsense reasoning and inductive biases to determine the role of soft causal relations on perceptions of story quality. We choose two story corpora in different genres: real-world mystery stories such as Sherlock Holmes—known for generally being consistent with everyday commonsense norms, and children's fairy tales such as Hansel and Gretel—stories which usually shatter commonsense expectations. Through these studies we further explore the broader issue of how the change in commonsense norms across storytelling genres affects perceptions of story quality.

Background and Related Work

Narrative generation systems that use symbolic planning (Lebowitz 1987; Gervás et al. 2005; Porteous and Cavazza 2009; Riedl and Young 2010; Ware and Young 2011) explicitly ensure causal relations between actions via predicate calculus operations over explicitly modeled action preconditions and post-conditions. These symbolic proposition

represent *hard* causal relations.

Neural language-model based approaches to story generation have typically overlooked causality or assumed it would emerge in the hidden state of neural networks. Roemmele and Gordon (2018) use LSTMs with skip-thought vector embeddings (Kiros et al. 2015) to generate stories. Similarly Clark, Ji, and Smith (2018) Martin et al. (2017, 2018) introduce semantic event abstractions known as events and decompose storytelling into the problems of generating event sequence and elaborating the events into natural language. Tambwekar et al. (2019) extends this work by fine-tuning language models to achieve a given goal, though goals are not necessarily achieved in a causality-preserving way as in symbolic planning. Fan, Lewis, and Dauphin (2018) and Ammanabrolu et al. (2020b) pursue hierarchical approaches to story generation, wherein a prompt is first generated and then transformed into a text passage. Yao et al. (2019) break down the problem of story generation into that of planning out a story and then generating from it.

Ippolito et al. (2019) look at filling in missing parts from a story by conditioning a text generator on rare words, also attempting to achieve balance between novelty and coherence. Donahue, Lee, and Liang (2020) also attempt to model storytelling along these lines, training a language model to fill in the blanks given left and right contexts. None of these methods explicitly incorporate commonsense knowledge into story generation.

An alternative machine learning based approach to story generation introduced by Li et al. (2013) is to first learn a *plot graph* that can then be used as a constrained search space for a sequence of story events. Plot graphs are directed acyclic dependency graphs where each node represents a plot point or event and the arcs between nodes represent *temporal* constraints. Inspired by this approach, we also attempt to learn a branching story graph structure that can be searched; however, instead of learning the plot graphs from a crowdsourced text corpus, we construct this graph by extracting commonsense inferences about causally related events.

Approaches to automated story generation that incorporate commonsense resources include the following. Rashkin et al. (2018) present an annotation framework specifically designed to examine the mental states of characters in commonsense based stories. Guan, Wang, and Huang (2019) incorporate external commonsense knowledge sources to explicitly improve story ending generation and Mao et al. (2019); Guan et al. (2020) look at fine tuning pre-trained transformer based language models (Vaswani et al. 2017) on commonsense sources like ConceptNet (Speer and Havasi 2012) and the BookCorpus (Kiros et al. 2015). These works, however, focus on improving what they call logicality and grammaticality, translating largely to local coherence, as opposed to analyzing perceptions of causality or overall story quality.

Soft Causal Relations

A *hard causal relation* implies that some world state transitions that are illegal—e.g., a character John cannot shoot Xavier if John is not in possession of a gun and the two characters are physically co-located. In contrast, a *soft causal relation* is mediated by the assumed reader's beliefs. Soft causality is therefore causality–normally a logical construct in narrative–mediated by the beliefs of the reader. It provides a causal ordering of events from the perspective of the reader instead of from the perspective of the author (whether human or agent). That is, a soft causal relation is a reasonable expectation of two non-mutually exclusive criteria: (a) certain activities are needed to achieve a character's goal, and (b) certain activities are in pursuit of future goals. The first clause draws on the psychological theory of the role of causality in story understanding by Trabasso and van den Broek (1985): readers attempt to understand "why" events occur by tracking causal relations as *enablement*—some event y cannot occur unless some preceding event x occurred. The second clause draws upon a theory of the role of character goal hierarchies in story understanding by Graesser, Lang, and Roberts (1991): readers attempt to understand "why" things happen by tracking and predicting character goal hierarchies. In both cases, whether an inference is made by reader is strongly dependent on what the reader's beliefs about the world are. In short, the key difference between hard and soft causality is the idea of expectations of causality via commonsense reasoning.

Commonsense knowledge is the set of commonly shared knowledge about how the world works. It enables us to form expectations about what will happen if we take certain courses of action and to infer things that likely happened in the past. Commonsense reasoning is the application of commonsense knowledge to specific contexts. Relevant to our work, commonsense reasoning might be applied to make inferences about what might have needed to have taken place for a character to arrive at a certain state—soft enablement—and what a reasonable next action would be based on what has happened so far—soft goal hierarchies.

Specifically for this paper, we use COMET (Bosselut et al. 2019) to model an assumed reader's commonsense knowledge. COMET is a transformer-based language model designed for commonsense inference and is trained on ATOMIC (Sap et al. 2019). ATOMIC is a dataset containing 877k instances of information relevant for everyday commonsense reasoning in the form of typed if-then relations with variables. ATOMIC is organized into different relation types such as "needs", "wants", "attributes", and "effects". We specifically use the relations for "wants" and "needs". An example of a cause using the *wants* relation is as follows, "if X tried to get away, then X *wants* to be free." Likewise, an example of an effect using the *needs* relation is, "if X scaled the wall, then X *needs* to know how to scale the wall."

The key difference between hard and soft causality is the idea of expectations of causality via commonsense reasoning and can be illustrated using the relations seen here. A hard causal relation requires verification and satisfaction of propositions, as in the example given in the paper - John cannot shoot Xavier if John is not in possession of a gun or they are not co-located. A soft causal relation here would be that the reader's belief that John dislikes Xavier and wants to fight him and thus as a result, he *wants* a weapon. Guns are weapons and thus there is a probability that John *needs* a weapon to fight Xavier.

In the next section we detail how we use the theory of soft causal relations, and COMET commonsense inferences

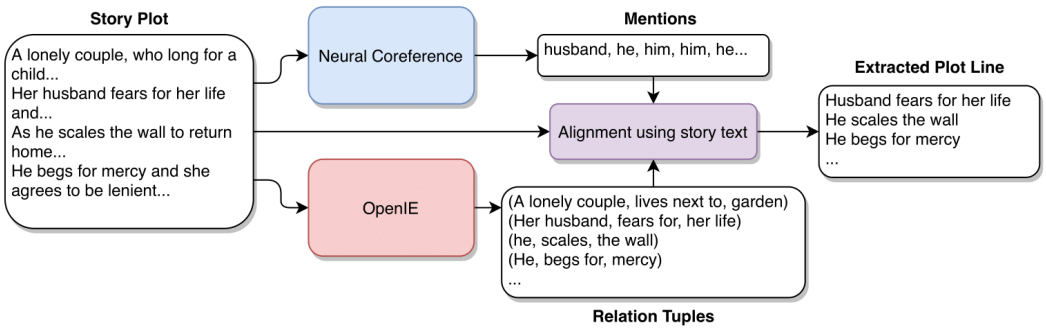


Figure 1: An illustration of high level plot point extraction.

about needs and wants, to generate stories. In section 5, we present the results of a human participant study that uses an evaluation of several systems in two distinct genres to probe how soft causal relations affect participant perceptions of story quality and coherence.

C2PO

This section presents the overall layout of C2PO. C2PO works by first extracting a set of high level plot points from a given textual story plot S and then generating a branching set of events that go between each high level plot point. The final story is obtained by walking the overall plot graph generated by joining each generated sub-graph.

Plot Extraction

The overall plot extraction process is described in Figure 1. In order to facilitate plot extraction, we propose a method that uses coreference resolution and information extraction to identify a set of plot points following a single character. First, we extract all the coreference clusters using a pre-trained neural coreference resolution model (Clark and Manning 2016). There can be multiple such clusters, each of which contains all mentions in the story belonging to a single possible character. We pick one of these clusters at random. Let $M = \{m_1, m_2, ..., m_n\}$ denote this cluster. Simultaneously, we also extract a set $\mathcal{R}$ of $\langle subject, relation, object \rangle$ triples from the story text using OpenIE (Angeli et al. 2015).

Once we have both of the set of mentions for a character and the triples for the story, we align them, attempting to find the subset of triples $\mathcal{P} \subset \mathcal{R}$ that are relevant for a single character on the basis of their character-level positions within the original story text. Both the neural coreference model and OpenIE are information retrieval systems and so we can identify the character-level offset or position of the retrieved information in the original story text. Let $pos(\cdot)$ be a function that can do this. The set of plot points is $\mathcal{P} = \{\langle s, r, o \rangle : pos(m) = pos(s), \forall m \in M, \langle s, r, o \rangle \in G\}$. The result is a sequence of relational tuples in which the character is the primary subject of the triple, ordered by when they first appeared in the original story text. Joining each triple together yields a subject-relation-object phrase which we refer to as a *plot point*.

Plot Graph Generation

Once we have established a series of plot points $\mathcal{P} = \{p_1, p_2, ..., p_n\}$, we move on to plot graph generation as illustrated in Figure 2. A plot graph is generated for each pair of adjacent plot points $(p_i, p_{i+1}), i \in \{1, .., n-1\}$ and then linked together in the order the plot points first appear in P to form a plot graph for an entire story.

The process to generate a plot graph between adjacent plot points p_1, p_2 is as follows. Starting from p_1, we use COMET (Bosselut et al. 2019) to generate candidate next events in the story. The *wants* relation indicates a direct forward cause—a character has a want and therefore performs an action. We recursively query COMET to generate k event candidates n times going forward starting with p_1; let this be $\mathcal{G}^f$. The *needs* relation indicates backward enablement—a character needed something to be true to do an action. We recursively query COMET to generate k event candidate n times going backward from p_2; let this be $\mathcal{G}^b$. This gives us two directed acyclic graphs as seen in Figure 2. The relations in $\mathcal{G}^f$ and $\mathcal{G}^b$ are weighted proportional to the likelihood score produced by COMET for each inference.

The next step is to look for the optimal way to link $\mathcal{G}^f$ and $\mathcal{G}^b$ and computing the probability of reaching a node $u \in \mathcal{G}^f$ looking at all nodes $\forall v \in \mathcal{G}^b$. Let $Pr^{needs}(u|v)$ be the probability of generating event e_2 as determined by COMET under the $needs$ relation, conditioned on e_1, and $Pr^{wants}(v|u)$ be the same but under the $wants$ relation. We define this link's weight as:

$$w(u, v) = \frac{Pr^{wants}(u|v)}{\alpha_u^{wants}} + \frac{Pr^{needs}(v|u)}{\alpha_v^{needs}} \quad (1)$$

where α_u^{wants} and α_v^{needs} are normalization constants. Here we set them equal to the probability of generating the word "to", a word in ATOMIC common to both relation types. This process is repeated for all nodes until we have found a set of optimal links. [2]

Finally, we link together the plot graphs for the entire sequence of plot points: $\mathcal{G} = \bigcup_{p_1, p_2} (\mathcal{G}^f_{p_1} \cup \mathcal{G}^b_{p_2}), \forall p_1, p_2 \in \mathcal{P}$

[2]COMET and ATOMIC can be replaced by any model designed for automated knowledge base completion and corresponding commonsense reasoning knowledge base by selecting the appropriate relations in the replacements.

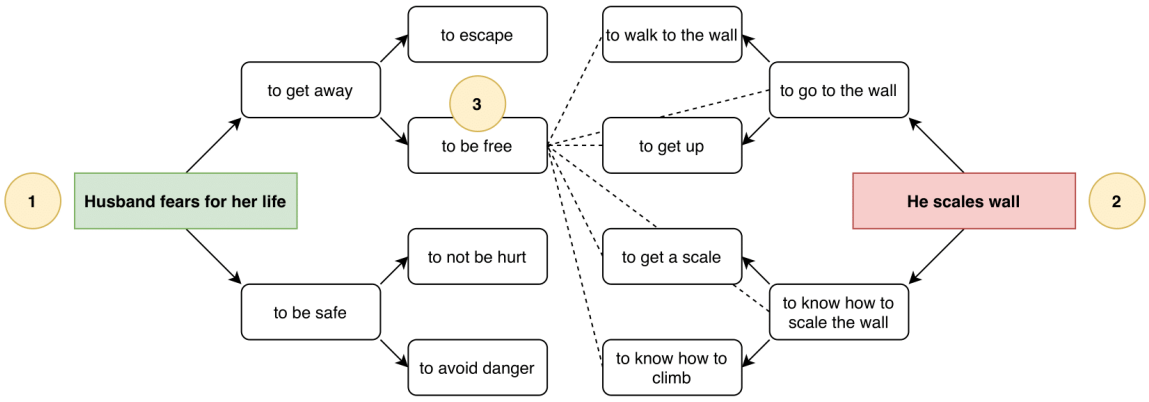


Figure 2: A demonstration of the plot graph generation process. 1 and 2 respectively indicate adjacent, extracted plot points. Dotted lines represent the process of finding the optimal link between the backward plot graph and node 3.

	Mystery	Fairy Tale
No. Stories	569	695
Sentences per story	23.36	24.80
Vocabulary size	21,238	16,452

Table 1: Dataset statistics.

	Commonsense	Storytelling
C2PO	✓	✓
BERT+infill	✓	
Hier. Fusion		✓

Table 2: Inductive biases of each system.

where p_1, p_2 are adjacent in $\mathcal{P}$. A story can be generated via a random walk of the graph from the first plot point p_1 to the last p_n. All random walks are guaranteed to terminate in p_n because $\mathcal{G}^b_{p_n}$ is constructed by branching backward from p_n. Likewise, each intermediate plot point $p_2 ... p_{n-1}$ is a node in $\mathcal{G}$ that all walks must pass through.

Experiments

We evaluate on a story dataset with two genres—mystery stories and fairy tales—first introduced by Ammanabrolu et al. (2020a)[3], statistics for the dataset can be found in Table 1. The data is partitioned into train and test splits in a 8:2 ratio, and the train split used to train C2PO and two baseline models (described below). A random set of 10 stories is chosen from each genre in the test set and high level plot points are extracted as described in the section on C2PO. For each model and for each set of high-level plot points and for each genre we generate three distinct stories for a total of $(3 \times 10 \times 2 \times 3 = 180$ stories. We generate three stories for each combination of model, plot point set, and genre to account for variance in stories that can be produced by the same high level plot due to the branching nature of C2PO as well as variance in the baselines' outputs. Standard automated language generation metrics such as perplexity and BLEU (Papineni et al. 2002) are known to be unreliable for creative generation tasks (Ammanabrolu et al. 2020b). The stories are thus evaluated using a human participant study, described below.

Baselines

We choose two baselines on the basis of the comparisons they afford (summarized in Table 2). Both are designed to perform text infilling tasks but differ based in their inductive biases. "Inductive biases" here specifically refer to a system's ability to model commonsense knowledge and if they were originally designed for storytelling or not.

BERT+infill The first baseline is a BERT (Devlin et al. 2018) based model that has not strictly been designed for storytelling (though BERT is trained on a corpus that includes story texts) and then adapted to perform text infilling. Although large-scale pre-trained language models are known not to be great storytellers, mostly due to them being unable to stay on track for any extended period of time (See et al. 2019), they have demonstrated knowledge of factual commonsense information by virtue of the amount of data they have been trained on (Petroni et al. 2019). Our problem setting requires us to generate a section of text between two consecutive high level plot points at a time, reminiscent of approaches taken by Ippolito et al. (2019) and Donahue, Lee, and Liang (2020) that condition a language model on left and right contexts to fill in blanks in a story. We follow a similar setup for this baseline, using BERT (Devlin et al. 2018) conditioned to attend to both previous tokens—the preceding plot point—and future tokens—the following plot point—to generate sequences (Lawrence, Kotnis, and Niepert 2019). BERT+infill is fine-tuned using this methodology on the high-level plot points extracted from our training data. Despite being similar to these prior methods, we note that BERT+infill utilizes no storytelling domain knowledge in its architecture and boils down to simple masked language modeling with multiple mask tokens.

Hierarchical Fusion Fan, Lewis, and Dauphin (2018) train their system—consisting of a convolutional sequence-to-sequence network with self-attention (Ott et al. 2019)—on

[3]https://github.com/rajammanabrolu/WorldGeneration

	C2PO	BERT+infill	Hierarchical Fusion
Mystery	**Holmes decides go.** Holmes wants to go. Holmes begins to see something. Holmes begins to look around. **Holmes notices a pair of trouser knees.** Holmes wants to clean up. Holmes begins take a shower. Holmes wants to get ready. Holmes wants to walk to the store. **Holmes taps in front of Wilson's shop.**	**Holmes decides go.** Holmes new friend initially stays. Holmes new son accepts goes. Holmes mother also stays. **Holmes notices a pair of trouser knees.** Holmes himself still watches. Holmes again is house ghost watches. Holmes insists he took watch. **Holmes taps in front of Wilson's shop.**	**Holmes decides go.** The room was silent. The room was silent. The air was heavy , and the room was quiet. **Holmes notices a pair of trouser knees.** The young man wasn't going to be a father. His parents weren't supposed to be a father. They had the best kids in the entire world. **Holmes taps in front of Wilson's shop.**
Fairy Tale	**Queen asks her mirror.** Queen wants to look better. Queen wants to try on clothes. Queen starts to be mad. **Queen is furious.** Queen tries to relax. Queen wants to take a nap. Queen starts to get up. Queen begins to approach someone. **She appears at a dwarfs'.** Queen starts to surprise everyone. Queen starts to have a party. queen wants to have money. Queen tries to buy poison comb. **She brushes with poisoned comb.**	**Queen asks her mirror.** Queen is still half smiles. Queen who had had frowns. Queen has always asked. **Queen is furious.** Queen wife of mary then flees. Queen wife husband anna maria refuses. Queen mistress queen mistress wives demands. **She appears at a dwarfs'.** Queen queen rose meets princess. Queen sees fairies she crowns fairies. Queen rises with beauty. **She brushes with poisoned comb.**	**Queen asks her mirror.** "What the ..." "You know I have no idea how I got here. You know I can't do anything about it." "I know I can't do anything about it." **Queen is furious.** A large, creature sits in the middle of a room with an odd looking cat on it. The creature is a strange looking cat, though it looks like the same cat is in its own room. Its fur is like a large, white slept. **She appears at a dwarfs'.** "So, you 're here to kill me," asked the man in the suit, with a slight hint of worry in his due. "Yes," replied the man in the suit. **She brushes with poisoned comb.**

Table 3: Examples of a story generated by each model in both genres given the same initial set of *bolded* high level plot points. Further randomly selected examples can be found in the Appendix .

	C2PO vs. BERT+	C2PO vs. Hier.	Tot.
Mystery	82	89	171
Fairy Tale	90	90	180
Total	172	179	351

Table 4: Participant count statistics.

the Reddit Writing Prompt corpus, where human-contributed prompts are paired with human-contributed stories. The system learns to first generate a prompt and then transform it into a story. This model's architecture is explicitly designed to tell stories and is suited for a type of storytelling wherein a prompt for a story is generated into a passage This type of training is particularly well suited to our setup of generating a story piece-by-piece using extracted high level plot points. We train the model from our training set using high level plots extracted from the stories, as described earlier, as the prompts and sections in between each of these extracted plot points as the story.

Human Evaluation Setup

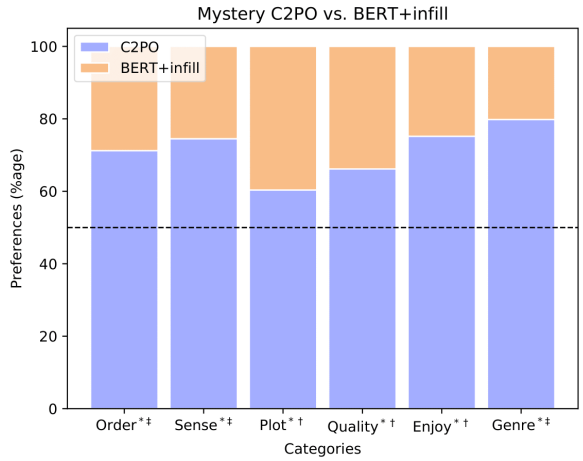
We have 10 sets of high level plots per genre and three generated stories per each plot for each of the models. We recruited 351 human participants via Mechanical Turk. Criteria for enrollment included: (a) fluency in English, and (b) demonstrating an understanding of commonsense based causality in stories. To screen participants for the latter we asked them to predict potential next events that could reasonably occur given a story scenario. An example of such a question asked can be found in the Appendix .

Human participants are given one story generated by C2PO and another evenly randomly picked from those generated by either BERT+infill or Hierarchical Fusion for the same plot. The order that these stories are presented in is randomized to account for bias induced due to the ordering effect (Olson and Kellogg 2014). Each story pairing is seen by at least three participants. Participant count statistics are given in Table 4.

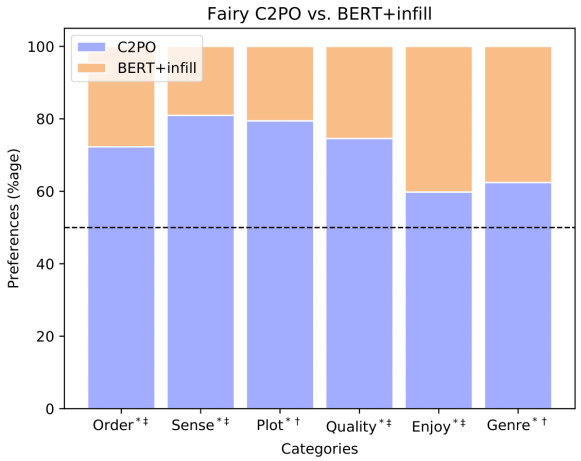
Participants are then asked a series of questions, each measuring a particular aspect of perceived story quality, comparing the C2PO generated model to one of the baselines. For each question they are asked to note down which story they preferred. The questions we use are adapted from Purdy et al. (2018) and have been used in multiple storytelling works as an indication of story quality (Tambwekar et al. 2019; Ammanabrolu et al. 2020b). Specifically, we ask:

- Which story's events occur in a more PLAUSIBLE ORDER?: as a proxy to indicate perceptions of overall causality within the story.
- Which story's sentences MAKE MORE SENSE given sentences before and after them?: to examine perceptions of local causality and commonsense reasoning in the story.
- Which story better follows a SINGLE PLOT?: for insight into perceptions of global coherence for the entire story.
- Which story is of HIGHER QUALITY?: as a measure of overall perceived story quality.
- Which story is more ENJOYABLE?: indicates story value.
- Which story better FITS A GENRE?: as a measure of how well the story matches commonsense knowledge specific to a genre, capturing the differences between the two genres.

For each of these questions, within a pairwise comparison, we perform a paired Mann-Whitney U test to assess statistical significance and additionally calculate Fleiss' κ (Kappa) value to measure inter-rater reliability.



(a) C2PO vs BERT+infill in the mystery genre.



(b) C2PO vs BERT+infill in the fairy genre.

Figure 3: Human evaluation results comparing C2PO vs BERT+infill. * indicates $p < 0.05$, ‡ indicates $\kappa > 0.4$ or moderate agreement, † indicates $\kappa > 0.2$ or fair agreement

	C2PO		**BERT+infill**		**Hierarchical**	
	Myst,	Fairy	Myst,	Fairy	Myst,	Fairy
Sent/Story	29.23	30.2	25.4	26.0	31.3	41.0
Words/Sent	4.94	5.04	4.62	4.79	7.21	5.75
Bigrams	312	317	356	357	380	402
Trigrams	1245	1353	1856	1870	2187	2190

Table 5: Statistics for generated stories. n-gram counts show unique ones measured with respect to those found in the test set of the initial story data.

Results and Analysis

There are a few dimensions along which we will attempt to analyze these results: (1) the inherent inductive biases of each model as seen in Table 2, (2) the two genres, and (3) the questions asked of the participants. The analysis will be performed hierarchically in the order just presented. Table 5 provides statistics on generated stories and Table 3 displays select examples of generated stories for each of the models in both genres.

C2PO vs BERT+infill

Figures 3a and 3b show the percentages that participants preferred C2PO versus the BERT+infill system for each dimension and for each story genre. C2PO is preferred over BERT+infill in both genres and in all dimensions. All of these results are statistically significant ($p < 0.05$) with fair-to-moderate inter-rater reliabilities.

For the mystery genre the greatest differences in preferences are observed with respect to enjoyability and genre resemblance. The systems were most similar with regard to their ability to maintain a single plot. For the fairy tale genre the greatest differences are seen in terms of the story events' plausible ordering, making sense causally, and the ability to maintain a single plot. The models were most similar with regard to their genre resemblance and enjoyability.

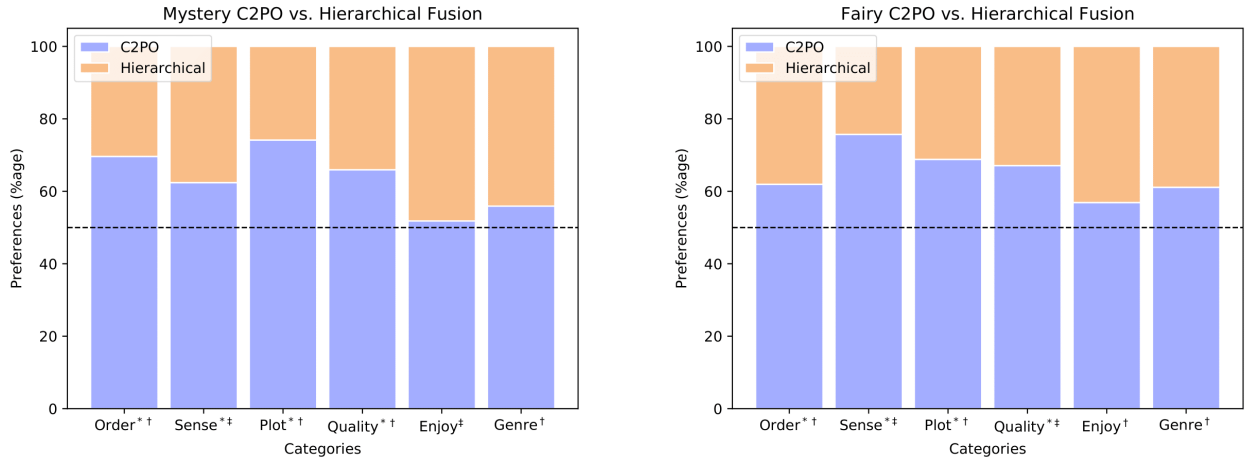
The questions that C2PO does particularly well on compared to BERT+infill are complementary across the genres. Enjoyability and genre resemblance are rated higher for C2PO in the mystery genre as opposed to fairy tales. We additionally observe that these two factors are highly, positively correlated using Spearman's Rank Order Correlation ($r_s = 0.56, p < 0.01$). Similarly, C2PO performed comparatively better in terms of plausible ordering, making sense causally, and the ability to maintain a single plot for fairy tales than for mysteries. These three factors are also highly, positively correlated with each other and in terms of overall perceived story quality ($0.6 > r_s > 0.55$, $p < 0.01$ for all pairwise comparisons).

This provides evidence that the brand of commonsense reasoning-based causality brought to bear by C2PO—needs and wants—works well in the mystery genre. The mystery genre follows everyday commonsense norms whereas the fairy tale genre is more likely to stray from commonsense norms. It can thus be inferred that genre-specific or thematic commonsense knowledge is required to improve perceptions of genre resemblance and enjoyability but does little in terms of metrics assessing local and global coherence in terms of causality.

C2PO vs Hierarchical Fusion

Figures 4a and 4b show the percentages of participants that preferred C2PO to Hierarchical Fusion. For the mystery genre, C2PO was preferred for the dimensions of plausible ordering, making causal sense, maintaining a single plot, and overall story quality. These dimensions were significantly different ($p < 0.05$). The dimensions of enjoyment and genre resemblance were not significantly different, meaning no system did better than the other.

We see a similar pattern for fairy tale stories: C2PO is preferred to hierarchical fusion for the same dimensions as



(a) C2PO vs. Hierarchical Fusion in the mystery genre.

(b) C2PO vs. Hierarchical Fusion in the fairy genre.

Figure 4: Human evaluation results comparing C2PO vs, Hierarchical Fusion. * indicates $p < 0.05$, ‡ indicates $\kappa > 0.4$ or moderate agreement, † indicates $\kappa > 0.2$ or fair agreement

the mystery genre and are not significantly different for enjoyment and genre resemblance.

Across genres, there is a positive correlation between metrics relating to coherence and overall perceived story quality ($0.6 > r_s > 0.5$, $p < 0.05$ for each pairwise comparison using Spearman's Rank Order Correlation). Also recall that the Hierarchical Fusion model contains an inductive bias for storytelling but does not model commonsense reasoning. This appears to indicate that genre resemblance and enjoyability are not dependant on causal, commonsense reasoning but rather on the how much the generated text "sounds like a story" but story quality still depends on overall coherence.

Broader Trends

There are two main trends that one can see across the models depending on their inductive biases (extent to which the models are trained for commonsense reasoning or storytelling). We observe these trends on the basis of the analysis presented so far as well as the examples of output stories found in Table 3. (1) Having commonsense reasoning abilities generally improves perceptions of local and global coherence in terms of causality with a caveat that what is perceived as commonsense can change across genres. When genre or domain specific commonsense knowledge matches "everyday" commonsense, it makes for an automated storyteller that is significantly more causal in nature. (2) Just commonsense reasoning without any sort of storytelling inductive bias incorporated—such as with pre-trained and finetuned language models which themselves have no real penchant for storytelling—into a model's design doesn't help, however, in terms of enjoyability and genre resemblance. The performance of Hierarchical Fusion in terms of enjoyability and genre resemblance—and the examples seen in Table 3—appear to indicate that models designed for storytelling do a better job of maintaining the writing style of a story but struggle with causality.

Conclusions

We intend for the findings of this work to be utilized by researchers studying automated storytelling, a standing AI grandchallenge requiring creative, long-form language generation. We explore the effects of *soft causal relations*—reasonable expectations by a reader regarding a story's progression—on human-based perceptions of overall story quality. We introduce C2PO as a way to use *soft causal relations* via transformer-based models trained for commonsense inference in storytelling.

A key insight from a human participant study, measuring a wide set of human perceived metrics, shows that the sum of the parts is indeed greater than the whole. Automated storytellers require both domain specific commonsense reasoning abilities as well as a storytelling inductive bias incorporated into the design of the system to perform well in terms of: local and global coherence on the basis of causality, enjoyability, genre resemblance, and overall story quality. Further, perceptions of causal, commonsense conforming coherence are highly correlated with overall story quality. We encourage authors of future work to build on these findings and more closely explore lines of research that use thematically relevant *soft causal relations* to improve automated storytellers.

Broader Impact

As an AI grandchallenge, automated storytelling consists of multiple sub-problems and thus has implications extending to the creation of learning agents that communicate effectively using natural language. Our system faces the same potential pitfalls as other contemporary language learning systems. Namely, it is prone to echoing the biases present in the dataset (Sheng et al. 2019). Some of these biases are thematic and give the reader a sense of the genre. *Prejudicial* biases are potentially more harmful. Story generation can involve non-normative language use, describing situations that

fictional characters engage in that would be considered inappropriate if enacted in the real world. Crowdsourcing—in the case of ATOMIC (Sap et al. 2019) which COMET (Bosselut et al. 2019) is trained on—and data curation—in the case of the story dataset used from Ammanabrolu et al. (2020a)—can mitigate, but do not entirely eliminate these biases. Story generation is broadly applicable (though not in its current state), from video game quests to conversational agents to book and movie generation. As with any broad capability technology, it can be put to purposes that are benign, malicious, or negligent. Fictional stories that are intentionally or unintentionally presented as true is a form of deception; we advise that future application developers using story generation technologies are very clear about the provenance of the story content delivered to an audience.

References

Ammanabrolu, P.; Cheung, W.; Tu, D.; Broniec, W.; and Riedl, M. O. 2020a. Bringing Stories Alive: Generating Interactive Fiction Worlds. In *16th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-20)*. URL http://arxiv.org/abs/2001.10161.

Ammanabrolu, P.; Tien, E.; Cheung, W.; Luo, Z.; Ma, W.; Martin, L. J.; and Riedl, M. O. 2020b. Story Realization: Expanding Plot Events into Sentences. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Angeli, G.; Premkumar, J.; Jose, M.; and Manning, C. D. 2015. Leveraging Linguistic Structure For Open Domain Information Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Çelikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Clark, E.; Ji, Y.; and Smith, N. A. 2018. Neural Text Generation in Stories Using Entity Representations as Context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2250–2260. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1204. URL https://www.aclweb.org/anthology/N18-1204.

Clark, K.; and Manning, C. D. 2016. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *EMNLP*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* .

Donahue, C.; Lee, M.; and Liang, P. 2020. Enabling Language Models to Fill in the Blanks. In *Association for Computational Linguistics*.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 889–898. URL https://arxiv.org/pdf/1805.04833.pdf.

Gervás, P.; Díaz-Agudo, B.; Peinado, F.; and Hervás, R. 2005. Story plot generation based on CBR. *Knowledge-Based Systems* 18(4-5): 235–242.

Graesser, A.; Lang, K. L.; and Roberts, R. M. 1991. Question Answering in the Context of Stories. *Journal of Experimental Psychology: General* 120(3): 254–277.

Guan, J.; Huang, F.; Zhao, Z.; Zhu, X.; and Huang, M. 2020. A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation. *Transactions of the Association for Computational Linguistics* .

Guan, J.; Wang, Y.; and Huang, M. 2019. Story Ending Generation with Incremental Encoding and Commonsense Knowledge. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*. URL https://arxiv.org/pdf/1808.10113.pdf.

Ippolito, D.; Grangier, D.; Callison-Burch, C.; and Eck, D. 2019. Unsupervised Hierarchical Story Infilling. In *Proceedings of the First Workshop on Narrative Understanding*, 37–43. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/W19-2405. URL https://www.aclweb.org/anthology/W19-2405.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, 3294–3302.

Lawrence, C.; Kotnis, B.; and Niepert, M. 2019. Attending to Future Tokens for Bidirectional Sequence Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1–10. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1001. URL https://www.aclweb.org/anthology/D19-1001.

Lebowitz, M. 1987. Planning Stories. In *Proceedings of the 9th Annual Conference of the Cogntive Science Society*, 234–242.

Li, B.; Lee-Urban, S.; Johnston, G.; and Riedl, M. O. 2013. Story Generation with Crowdsourced Plot Graphs. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, 598–604. AAAI Press. URL http://dl.acm.org/citation.cfm?id=2891460.2891543.

Mao, H. H.; Majumder, B. P.; McAuley, J.; and Cottrell, G. 2019. Improving Neural Story Generation by Targeted Common Sense Grounding. 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)*, 5988–5993. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1615. URL https://www.aclweb.org/anthology/D19-1615.

Martin, L. J.; Ammanabrolu, P.; Wang, X.; Hancock, W.; Singh, S.; Harrison, B.; and Riedl, M. O. 2018. Event Representations for Automated Story Generation with Deep Neural Nets. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 868–875. New Orleans, Louisiana.

Martin, L. J.; Ammanabrolu, P.; Wang, X.; Singh, S.; Harrison, B.; Dhuliawala, M.; Tambwekar, P.; Mehta, A.; Arora, R.; Dass, N.; Purdy, C.; and Riedl, M. O. 2017. Improvisational Storytelling Agents. In *Workshop on Machine Learning for Creativity and Design (NeurIPS 2017)*. Long Beach, CA. URL https://nips2017creativity.github.io/doc/Improvisational{_}Agents.pdf.

Olson, J. S.; and Kellogg, W. A. 2014. *Ways of Knowing in HCI*. Springer Publishing Company, Incorporated. ISBN 1493903772.

Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; and Auli, M. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. doi:10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1250. URL https://www.aclweb.org/anthology/D19-1250.

Porteous, J.; and Cavazza, M. 2009. Controlling narrative generation with planning trajectories: The role of constraints. In *Joint International Conference on Interactive Digital Storytelling*, volume 5915 LNCS, 234–245. Springer. ISBN 3642106420. ISSN 03029743.

Purdy, C.; Wang, X.; He, L.; and Riedl, M. 2018. Predicting Generated Story Quality with Quantitative Measures. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*. URL https://aaai.org/ocs/index.php/AIIDE/AIIDE18/paper/view/18106.

Rashkin, H.; Bosselut, A.; Sap, M.; Knight, K.; and Choi, Y. 2018. Modeling Naive Psychology of Characters in Simple Commonsense Stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2289–2299. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/P18-1213. URL https://www.aclweb.org/anthology/P18-1213.

Riedl, M. O.; and Young, R. M. 2010. Narrative Planning: Balancing Plot and Character. *Journal of Artificial Intelligence Research* 39: 217–267. URL https://www.cc.gatech.edu/{~}riedl/pubs/jair.pdf.

Roemmele, M.; and Gordon, A. S. 2018. An Encoder-decoder Approach to Predicting Causal Relations in Stories. In *Proceedings of the First Workshop on Storytelling*, 50–59. New Orleans, Louisiana: Association for Computational Linguistics. URL http://aclweb.org/anthology/W18-1506http://people.ict.usc.edu/{~}gordon/publications/NAACL-WS18A.PDF.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3027–3035.

See, A.; Pappu, A.; Saxena, R.; Yerukola, A.; and Manning, C. D. 2019. Do Massively Pretrained Language Models Make Better Storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 843–861. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/K19-1079. URL https://www.aclweb.org/anthology/K19-1079.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 3407–3412. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1339. URL https://www.aclweb.org/anthology/D19-1339.

Speer, R.; and Havasi, C. 2012. Representing General Relational Knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. ISBN 978-2-9517408-7-7.

Tambwekar, P.; Dhuliawala, M.; Martin, L. J.; Mehta, A.; Harrison, B.; and Riedl, M. O. 2019. Controllable Neural Story Plot Generation via Reward Shaping. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. URL https://www.ijcai.org/proceedings/2019/829.

Trabasso, T.; and van den Broek, P. 1985. Causal Thinking and the Representation of Narrative Events. *Journal of Memory and Language* 24: 612–630.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Ware, S.; and Young, R. M. 2011. CPOCL: A narrative planner supporting conflict. In *Proceedings of the 7th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Yao, L.; Peng, N.; Weischedel, R.; Knight, K.; Zhao, D.; and Yan, R. 2019. Plan-And-Write: Towards Better Automatic Storytelling. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*.