Back to the Future: Unsupervised Backprop-based Decoding for Counterfactual and Abductive Commonsense Reasoning

Lianhui Qin[†‡]   **Vered Shwartz** [†‡]   **Peter West** [†‡]   **Chandra Bhagavatula**[‡]
Jena D. Hwang [‡]   **Ronan Le Bras** [‡]   **Antoine Bosselut** [†‡]   **Yejin Choi**[†‡]

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
{lianhuiq, pawest, antoineb, yejin}@cs.washington.edu
{vered, chandrab, jenah, ronanlb}@allenai.org

Abstract

Abductive and counterfactual reasoning, core abilities of everyday human cognition, require reasoning about what might have happened at time t, while conditioning on multiple contexts from the relative past and future. However, simultaneous incorporation of past and future contexts using generative language models (LMs) can be challenging, as they are trained either to condition only on the past context or to perform narrowly scoped text-infilling.

In this paper, we propose DELOREAN, a new unsupervised decoding algorithm that can flexibly incorporate both the past and future contexts using only off-the-shelf, left-to-right language models and no supervision. The key intuition of our algorithm is incorporating the future through *back-propagation*, during which, we only update the internal representation of the output while fixing the model parameters. By alternating between forward and backward propagation, DELOREAN can decode the output representation that reflects both the left and right contexts. We demonstrate that our approach is general and applicable to two nonmonotonic reasoning tasks: abductive text generation and counterfactual story revision, where DELOREAN outperforms a range of unsupervised and some supervised methods, based on automatic and human evaluation.[1]

1 Introduction

Everyday causal reasoning requires reasoning about the likely explanations to partially observable past and future (*abductive* reasoning (Peirce, 1960)) and reasoning about the alternative future based on counterfactual past (*counterfactual* reasoning). Such *nonmonotonic* reasoning requires

---

[1]Code is available at https://github.com/qkaren/unsup_gen_for_cms_reasoning
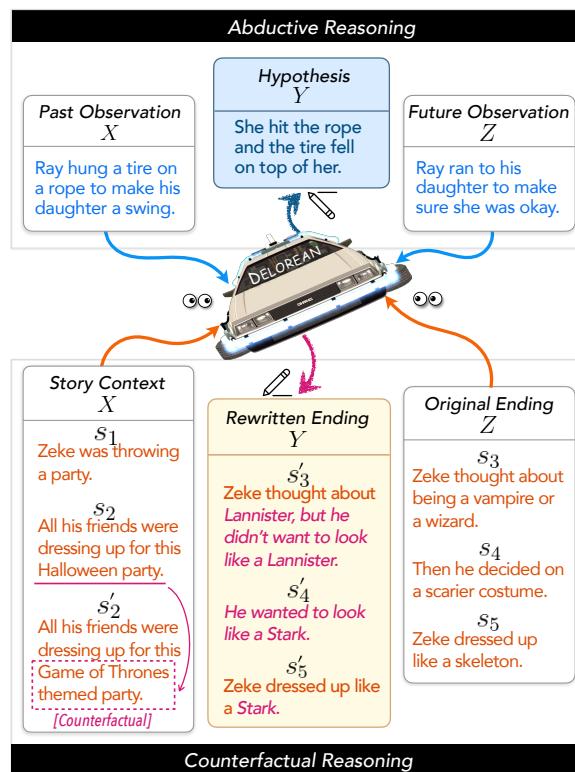

Figure 1: DELOREAN, our proposed method, with generated reasoning results. **Top**: the goal in abductive reasoning is to generate a hypothesis (Y) of what happened between the observed past (X) and future (Z) contexts. **Bottom**: In counterfactual reasoning, given a story context altered by a counterfactual condition, X, and the original ending Z, the goal is to generate a new ending Y which is coherent with X while remaining similar to Z. The story from TIMETRAVEL (Qin et al., 2019a) consists of five sentences. Our approach alternates forward (left-to-right) and backward (right-to-left) passes that iteratively refine the generated texts w.r.t context from each side.

inferring plausible but potentially defeasible conclusions from incomplete or hypothetical observations (Reiter, 1988). While humans are remarkably good at this type of causal reasoning, developing AI systems capable of nonmonotonic reasoning for

a wide range of situations describable in natural language has been a major open research question.

More concretely, with abductive reasoning, the goal is to find the most plausible explanation for incomplete observations (Peirce, 1960). In the top part of Figure 1, given the first observation that Ray is "making his daughter a swing" and the later observation that he "ran to [her] to make sure she was okay," we can hypothesize that she somehow got hurt by the swing.

In contrast, counterfactual reasoning concerns the causal changes to future events given a change in the past condition (i.e., "counterfactual condition"; Goodman, 1947). For example, the bottom part of Figure 1 shows the *original* five sentence story (S_1, ..., S_5) and an alternative *counterfactual condition* given in S_2'—that instead of being a generic "Halloween party", the new counterfactual condition is that it is going to be a "Game of Thrones themed party"! Given these, the problem we want to solve is to update the future events (S_3', ..., S_5'), so that instead of "Zeke dressed up as skeleton", we have "Zeke dressed up like a Stark".[2]

Recently, two tasks and corresponding benchmarks have been introduced to tackle language-based nonmonotonic reasoning: the $\mathcal{ART}$ dataset for abductive NLG (Bhagavatula et al., 2019), and the TIMETRAVEL dataset for counterfactual story rewriting (Qin et al., 2019a). Both tasks are framed as conditional generation, with multiple contexts to condition on. The currently dominant paradigm for conditional text generation tasks is fine-tuning pre-trained language models (LMs), such as GPT2 (Radford et al., 2019a), on large-scale training data for supervision. However, despite the large number of training examples, supervised approaches still perform considerably worse than humans and are subject to developing superficial strategies such as repeating the observations as is or memorizing prevalent surface patters specific in the dataset (Qin et al., 2019a). Furthermore, having to require large-scale training data for each domain and task would be utterly inefficient for broad-coverage nonmonotonic reasoning in language.

In this paper, we investigate an alternative path toward language-based nonmonotonic reasoning using pre-trained language models as is. Intuitively, both the abductive and counterfactual reasoning

requires learning coherent patterns in narrative, which should be already available in large-scale pretrained language models. However, the key challenge is that most generative language models are trained to condition only on the left context, or to perform narrowly scoped text-infilling.

This paper presents DELOREAN: DEcoding for nonmonotonic LOgical REAsoNing, an unsupervised decoding algorithm that only assumes off-the-shelf left-to-right language models with no supervision. The key intuition of our algorithm is incorporating the future through back-propagation, during which, we only update the internal representation of the output while fixing the model parameters. More specifically, DELOREAN alternates between the forward and backward passes, where the forward pass performs left-to-right inference given the left context (roughly maximizing $P(Y|X)$ in Figure 1), while the backward pass instills the right constraint through right-to-left backpropagation with a task-specific loss (roughly maximizing $P(Z|XY)$). The forward and backward outputs are mixed into a single vector, from which tokens are sampled to generate the desired output. To choose the best output across iterations, we employ an unsupervised ranking step based on BERT's next sentence prediction task to measure coherence (Devlin et al., 2018).

On both tasks, DELOREAN outperforms all other unsupervised methods in terms of both automatic metrics and human evaluation, demonstrating that nonmonotonic reasoning through conditional decoding is a promising research direction. Moreover, outputs produced by our model are judged as more coherent than those from the supervised models. In sum, our study shows that backpropagation-based decoding may enable additional future applications of unsupervised generation and reasoning.

2 Background

Most NLP benchmarks have focused on reasoning about information that is *entailed* from the premise. For instance, natural language inference (NLI; Bowman et al., 2015) focuses primarily on whether a hypothesis is entailed from a given premise, which means the information stated in the hypothesis is a subset of the information provided in the premise. However, it has been noted that human reasoning is often the other way, where hypotheses often contain new information that was not available in the premise, but plausibly true (but

---

[2]"Lannister" in S_3' and "Stark" in S_4' and S_5' refer to character names in the TV show, "Game of the Thrones." All the output text shown in Figure 1 is the actual system output from DELOREAN.
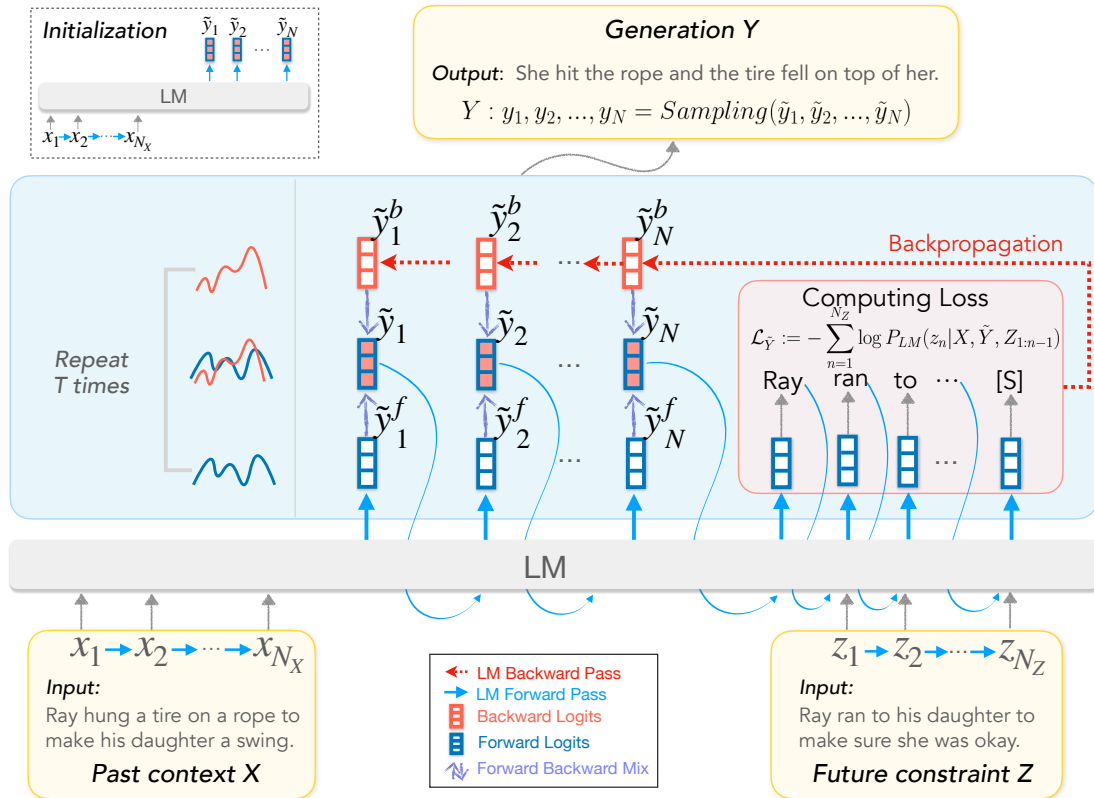
Figure 2: Illustration of the DELOREAN decoding procedure, using abductive reasoning as an example. At initialization (upper-left box), the language model (LM) initializes the logits ▯ $\tilde{Y} = \{\tilde{y}_1, \ldots, \tilde{y}_N\}$ of the hypothesis by reading the past context X and generating a continuation with regular decoding. At each forward-backward iteration, we compute the task-specific **loss** $\mathcal{L}_{\tilde{Y}}$ of the logits ▯ based on the future constraint Z (red box). The **backward pass** then performs back-propagation and produces the backward logits ▯ $\tilde{Y}^b = \{\tilde{y}_1^b, \ldots, \tilde{y}_N^b\}$. In the subsequent **forward pass**, for each step n, we compute the forward logits ▯ $\tilde{y}_n^f$ conditioning on the preceding logits ▯ $\tilde{y}_{n-1}$, and then mix it with the respective backward logits ▯ to produce the new logits ▯ $\tilde{y}_n$ at step n.

possibly defeasible with new additional context) (Johnson-Laird, 2006; Mercier and Sperber, 2017). This type of reasoning corresponds to nonmonotonic reasoning (Kraus et al., 1990), as it contradicts the monotonicity property according to which valid arguments cannot be made invalid by adding premises. We study two tasks of that nature: abductive reasoning (§2.1) and counterfactual reasoning (§2.2).

2.1 Abductive Reasoning

Abductive reasoning aims at finding the most likely explanation to partial observations (Peirce, 1960). It has a central role in the human ability to "read between the lines," and is crucial for language acquisition (Andersen, 1973), understanding sentences in discourse (Hobbs et al., 1993), and many more. Despite the importance, however, relatively little focus has been given to it in NLP research.

Recently, Bhagavatula et al. (2019) propose the abductive reasoning task. Given two observations, the goal is to determine the most likely explanation of what happened in-between. The dataset introduced for the task, $\mathcal{ART}$, consists of 20k observations derived from the first and last sentence of stories in the ROCStories dataset (Mostafazadeh et al., 2016a). We focus on the abductive NLG setup introduced in the paper, which is framed as a conditional generation task where a plausible explanation to the observations must be generated using language. The authors reported the performance of several pre-trained LM-based baselines and showed promises and limitations of such approaches.

2.2 Counterfactual Reasoning

Counterfactual reasoning aims at inferring alternative past events that could have happened given a certain change in conditions (Goodman, 1947; Starr, 2019). While counterfactual reasoning plays an important role in AI systems (Isard, 1974; Gins-

berg, 1986), it requires causal reasoning abilities, which are arguably absent from current association-based AI (Pearl and Mackenzie, 2018). While there has been work on counterfactual reasoning in NLP, including recognizing counterfactuals in text (Son et al., 2017), and improving the performance of NLP tasks using *counterfactual learning* (Lawrence et al., 2017; Lawrence and Riezler, 2018), it remains a major research challenge.

Recently, Qin et al. (2019a) introduce the task of counterfactual story generation. Given a 5-sentence original story, and an alternative context in which the second sentence of the story was altered by a counterfactual, the task is to generate a new 3-sentence story ending that addresses the alternative beginning while minimally editing the original ending. The associated TIMETRAVEL dataset is based on fictional narratives from ROCStories, for which counterfactual contexts and alternative endings are crowdsourced, yielding 29,849 problem instances. Qin et al. (2019a) report several baseline performances, and find that models based on pre-trained LMs produce output that recognize the counterfactual, but generated endings which deviated considerably from the original storyline. In contrast, in the supervised setup, models optimize the easier of the two goals and generate endings that are overly similar to the original endings.

3 The DELOREAN Approach

Humans make inferences based on available information and refine them when new information arrives. Since currently available pre-trained LMs generate text by sequentially predicting the next token from left to right, they are incapable of conditioning on future constraints. Therefore, we propose DELOREAN: an unsupervised backprop-based decoding algorithm, which is summarized in Algorithm 1, illustrated in Figure 2, and detailed below. DELOREAN intermittently refines the predictions to cohere with either the context or the constraints (Section 3.1). The candidate generations are then ranked by coherence (Section 3.2).

3.1 Decoding Strategy

Given context text X, the goal is to generate continuation text $Y = (y_1, \ldots, y_N)$, such that Y satisfies certain constraints according to the reasoning tasks, usually defined based on another context Z (see Figure 1; we discuss the task-specific constraints in the respective task sections).

Algorithm 1: DELOREAN Decoding

Input: Pre-trained language model (LM)
      Context X
      Future constraint Z
1: Initialize logits $\tilde{Y}^{(0)}$
2: Initialize Y_s, list of candidate generations
3: **for** $t \leftarrow 1$ to T **do**
4:     // Backward pass
5:     **for** $n \leftarrow N$ to 1 **do**
6:         Compute backward logits $\tilde{\boldsymbol{y}}_n^{\text{b}}$, Eq.(1)
7:     **end for**
8:     // Forward pass
9:     **for** $n \leftarrow 1$ to N **do**
10:        Compute forward logits $\tilde{\boldsymbol{y}}_n^{\text{f}}$, Eq.(2)
11:        Mix forward and backward logits, Eq.(3)
12:     **end for**
13:     Sample candidate Y from logits $\tilde{Y}$ and add to Y_s
14: **end for**
15: Rank Y_s by coherence
Output: The most coherent generated text Y from Y_s

The proposed approach interleaves two procedures, namely, *forward* and *backward*, that produce and iteratively refine the generation, for a predefined number of iterations T. In particular, the *forward* pass ensures the generated text is a fluent continuation of the context X, while the *backward* pass informs the model about the constraint and steers the generation to satisfy it.

As detailed below, the backward pass uses gradient descent to update the generation Y. However, Y is a discrete text that is not differentiable. Instead, throughout the algorithm, we maintain a soft representation of the sequence $\tilde{Y} = (\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_N)$, where $\tilde{\boldsymbol{y}}_n \in \mathbb{R}^V$ represents the logits of the n-th token and V is the vocabulary size. After the logits are refined over multiple iterations of the forward and backward passes, we generate discrete text at each step by sampling from $y_n \sim \text{softmax}(\tilde{\boldsymbol{y}}_n/\tau)$, where $\tau > 0$ is the temperature.

We start by initializing the logits before the first iteration, $\tilde{Y}^{(0)} = (\tilde{y}_1^{(0)} \ldots \tilde{y}_N^{(0)})$, by feeding the context X into the LM and greedily decoding N continuation tokens.

Backward The backward pass uses gradient backpropagation to update the generation with respect to the constraint. Specifically, we express the task-specific constraint as a loss function $\mathcal{L}(X, \tilde{Y}^{(t-1)}, Z)$ that evaluates how well the generation Y (approximated with the soft representation $\tilde{Y}$) obeys the constraint (see the subsequent sections for concrete instantiations of the loss). The goal of this pass is thus to minimize the loss w.r.t the generation. Specifically, at iteration t, for each

step n in the generation, we update its logits with:

$$\tilde{\boldsymbol{y}}_n^{(t),\mathrm{b}} = \tilde{\boldsymbol{y}}_n^{(t-1)} - \lambda \cdot \nabla_{\tilde{\boldsymbol{y}}_n} \mathcal{L}(X, \tilde{Y}^{(t-1)}, Z), \quad (1)$$

where $\nabla_{\tilde{\boldsymbol{y}}_n} \mathcal{L}(X, \tilde{Y}^{(t-1)}, Z)$ is the gradient of the constraint-informed loss $\mathcal{L}$ w.r.t the n-th logits, and $\lambda \in \mathbb{R}$ is the step size. In practice, we may repeat the gradient updates multiple times in a single pass.

Forward The forward pass ensures that Y is fluent and coherent with the preceding context X. At iteration t, for a particular step n, we compute the forward logits with the LM:

$$\tilde{\boldsymbol{y}}_n^{(t),\mathrm{f}} = \mathrm{LM}(X, \tilde{Y}_{1:n-1}^{(t)}). \quad (2)$$

We then *mix* the nth-step forward and backward logits to get the final logits of iteration t:

$$\tilde{\boldsymbol{y}}_n^{(t)} = \gamma \cdot \tilde{\boldsymbol{y}}_n^{(t),\mathrm{f}} + (1 - \gamma) \cdot \tilde{\boldsymbol{y}}_n^{(t),\mathrm{b}}, \quad (3)$$

where $0 < \gamma < 1$ is the mixing weight. The resulting logits $\tilde{\boldsymbol{y}}_n^{(t)}$ are then fed to the LM to compute the forward logits at the $(n+1)$th step (Eq.2). This way, information from the backward pass is integrated into the left-to-right generation process to produce text that is informed by the constraint.

We pre-define the number of tokens N required by the backward pass, but we allow the forward pass to generate more than N tokens if those are needed to obtain complete sentences. In that case, we set the logits of the extra tokens to the forward logits, without mixing: $\tilde{\boldsymbol{y}}_n^{(t)} = \tilde{\boldsymbol{y}}_n^{(t),\mathrm{f}}$ for $n > N$. We then prune any trailing tokens in the sampled text to get complete sentences.

3.2 Ranking

The output of the decoding step is a list of candidate generations for each iteration: $Y_s = \{Y^{(t)} | t = 1, ..., T\}$. We further use an unsupervised approach to rank and pick the best sample as the final output. Specifically, we take advantage of the BERT model, which was pre-trained with a next-sentence prediction (NSP) objective. Given two sentences A and B, we use NSP to compute the likelihood of B following A as a proxy for coherence:

$$c(A, B) = \mathrm{BERT\_NSP}(A, B), \quad (4)$$

where $c(\cdot, \cdot)$ denotes the coherence score. This score is used to evaluate the quality of a given candidate continuation Y by measuring (1) its compatibility with the subsequent text of the context X, (2) the internal consistency of Y if it consists of multiple sentences, and (3) the compatibility of Y with its right-side text when it is applicable.

Model	BLEU-4	ROUGE-L	BERT
Supervised			
Sup	32.82	25.60	49.38
+*COMET-Emb*	33.97	26.06	49.71
Unsupervised			
Zero-Shot$_X$	18.30	14.99	39.36
Zero-Shot$_{ZX}$	15.90	14.23	40.03
Zero-Shot$_X$-Ranked	19.24	16.76	41.58
Zero-Shot$_{ZX}$-Ranked	20.13	17.25	41.93
DELOREAN	**22.60**	**18.94**	**42.86**
Human	53.56	30.40	53.30

Table 1: Automatic evaluation results on the abductive task, using the test set of $\mathcal{ART}$.

4 Task 1: Abductive Reasoning

Each instance in the $\mathcal{ART}$ dataset consists of two observations O_1, O_2 and a hypothesis H that explains the two observations. These inputs naturally map to X, Z and Y in our framework. Formally, the abductive generation task aims to maximize $P(Y|X, Z)$ – i.e. models must consider both left and right contexts (X and Z) jointly.

4.1 Task Setup

Constraints We maximize Z given $X\tilde{Y}$ by defining the loss function as the cross-entropy loss of generating Z given $X\tilde{Y}$ with the LM:[3]

$$\mathcal{L}(X, \tilde{Y}, Z) := -\sum_{n=1}^{N_Z} \log P_{\mathrm{LM}}(z_n | X, \tilde{Y}, Z_{1:n-1}), \quad (5)$$

where $P_{\mathrm{LM}}(a_j | a_{1:j-1})$ is the likelihood of generating token a_j given the preceding text $a_{1:j-1}$.

Ranking We rank candidates by the overall coherence after inserting Y in between X and Z:

$$\mathrm{ranking\_score}(Y) = c(XY, Z) + c(X, YZ). \quad (6)$$

Hyperparameters We use GPT2-345M (Radford et al., 2019b) as the pre-trained LM for all models. We use the $\mathcal{ART}$ development set to select hyperparameters. We use greedy decoding for our method and top k decoding (Fan et al., 2018) ($k = 40, \tau = 0.7$) for our baselines. Other hyperparameters are outlined in Appendix A.1.

4.2 Experimental Setup

Baselines We compare our method against baselines from Bhagavatula et al. (2019). The unsupervised baselines use a pre-trained GPT-2 model

---

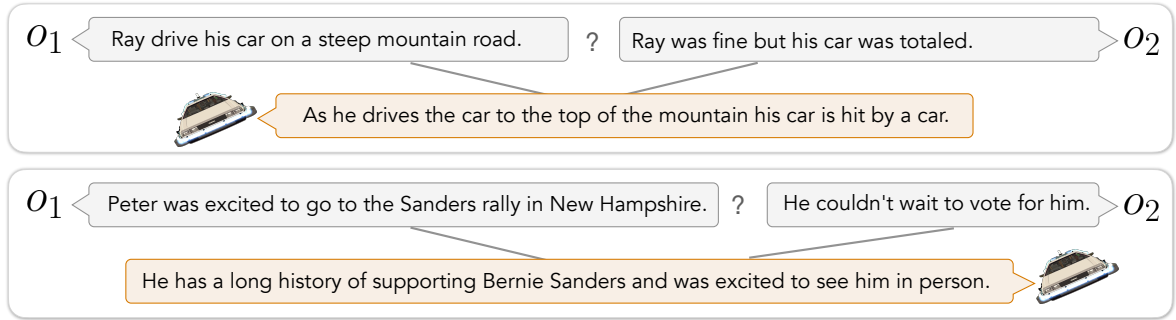[3]Note that this is applied to each prefix of $\tilde{Y}$, although some of them are not complete sentences.

Figure 3: Examples of generated hypotheses on three abductive reasoning cases. Given observations O1 and O2, DELOREAN generates a hypothesis explaining the observations.

to generate Y given a prompt text—either the observation X alone (Zero-Shot$_X$) or $Z\langle e\rangle X$ (Zero-Shot$_{ZX}$), where $\langle e\rangle$ denotes a special end-of-text token. The supervised method (Sup) follows the same input format as Zero-Shot$_{ZX}$, but fine-tunes GPT-2 on the $\mathcal{ART}$ training set. Finally, our knowledge-informed baseline (+*COMET-Emb*) further augments the representation of Sup with knowledge from COMET (Bosselut et al., 2019).

To separately study the contribution of our decoding strategy and ranking component, we also report the performance of ranking the baseline outputs. Specifically, we let each baseline generate 20 candidates and rank them by coherence (Eq. 6).[4]

4.3 Results

Automatic Evaluation We report the same metrics as Bhagavatula et al. (2019): BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004) and BERTSCORE (Zhang et al., 2019) (with the *bert-base-uncased* model). The results in Table 1 show that DELOREAN performs best among the unsupervised systems across all metrics. We also note that our ranking step improves both the performance of our model and that of the zero-shot baselines.

Human Evaluation We conduct two sets of human evaluations on 100 test examples using crowd-workers from Amazon Mechanical Turk. In the scoring setting, presented in Table 2, workers were presented a pair of observations (X and Z) and a generated hypothesis Y, and asked to rate the coherence of the hypothesis with respect to the observation X (X-Y), the observation Z (Y-Z), and both (X-Y-Z), on a 4-point Likert scale. In the

Model	X-Y	Y-Z	X-Y-Z
Supervised			
Sup	0.510	0.375	0.314
+*COMET-Emb*	0.466	0.342	0.286
Unsupervised			
Zero-Shot$_{ZX}$	0.233	0.103	0.108
Zero-Shot$_X$-Ranked	0.478	0.208	0.195
Zero-Shot$_{ZX}$-Ranked	0.474	0.238	0.236
DELOREAN	**0.522**	**0.325**	**0.297**
Human	0.879	0.823	0.783

Table 2: Human calibration results on test set of $\mathcal{ART}$. All scores are normalized to $[0, 1]$.

Overall - Human Judges Preferred				
Our model		Neutral		Comparator
DELOREAN	21%	43%	**36%**	Sup
DELOREAN	25%	44%	**31%**	+*COMET-Emb*
DELOREAN	**23%**	62%	15%	Zero-Shot$_X$-Ranked
DELOREAN	**27%**	50%	23%	Zero-Shot$_{XZ}$-Ranked
DELOREAN	3%	11%	**86%**	Human

Table 3: Human pairwise comparison results on the test set of $\mathcal{ART}$, between DELOREAN and each of the baselines, by jointly considering all 3 criteria from Table 2. "Neutral" means "equally good/bad".

pairwise comparison setting, presented in Table 3, workers were presented the outputs from a pair of systems (DELOREAN and baseline) and asked to choose the better output in terms of the same coherence criteria. Each example was labeled by 3 workers.[5]

In both evaluation setups, our method substantially outperform the unsupervised baselines, achieving a relative improvement of $36\% - 215\%$ with respect to Y-Z coherence. Our method also outperform the supervised methods with respect to X-Y coherence (Table 2), and achieve competitive performance in the pairwise comparison (Table 3).

---

[4]We tried ablating the ranking component from our method in preliminary experiments, and found that ranking is essential to obtaining good performance. By adding ranking to our baselines, we assess the contribution of our decoding strategy.

[5]The average inter-rater agreement measured by Fleiss' $\kappa = 0.44$ ("moderate agreement") (Fleiss, 1971).

	BLEU	ROUGE	BERT
Supervised + Discriminative			
Sup+Disc	75.71	72.72	62.39
Unsupervised+ Discriminative			
Recon+CF	**75.92**	**70.93**	62.49
Unsupervised			
FT	4.06	24.09	62.55
FT+CF	4.02	24.35	62.63
Pretrained-only			
Zero-Shot$_{s_1 s_2'}$	1.74	21.41	59.31
Zero-Shot$_{s_1 s_2'}$-Ranked	2.26	25.81	60.07
DELOREAN	21.35	40.73	**63.36**
Human	64.93	67.64	61.87

Table 4: Automatic evaluation results of counterfactual story rewriting, on the test set of TIMETRAVEL.

Coherence - Human Judges Preferred			
Our model	Neutral	Comparator	
DELOREAN **25%**	58%	17%	Sup+Disc
DELOREAN **23%**	70%	7%	Recon+CF
DELOREAN 22%	48%	**30%**	FT
DELOREAN 18%	60%	**22%**	Zero-Shot$_{s_1 s_2'}$
DELOREAN 27%	42%	**31%**	Zero-Shot$_{s_1 s_2'}$-Ranked
DELOREAN 10%	29%	**61%**	Human

Min-Edits - Human Judges Preferred			
Our model	Neutral	Comparator	
DELOREAN 4%	17%	**79%**	Sup+Disc
DELOREAN 1%	14%	**85%**	Recon+CF
DELOREAN **21%**	76%	3%	FT
DELOREAN **28%**	71%	1%	Zero-Shot$_{s_1 s_2'}$
DELOREAN **37%**	56%	7%	Zero-Shot$_{s_1 s_2'}$-Ranked
M+Sup 8%	22%	**70%**	Human

Table 5: Human pairwise comparison results on the counterfactual task, between our best model and each baseline with respect to coherence and min-edits.

Again, the ranking component contributes to increasing performance for the zero-shot baselines. Finally, the large performance gap between the methods and human-written explanations stresses the difficulty of this reasoning task and warrants future research.

Qualitative Analysis Figure 3 presents two example outputs produced by DELOREAN. We can see our approach generates reasonable hypotheses by taking into account both the past and future contexts. For instance, in the first example, the future observation (O2) "car was totaled" indicates that Ray had a car accident, which is correctly captured in the generated hypothesis "car is hit by a car".
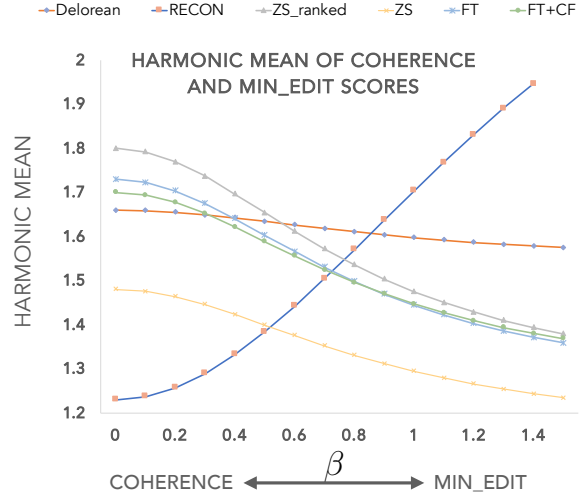

Figure 4: Human calibration results for counterfactual generation in terms of weighted harmonic mean of coherence and min-edit, $H_\beta = \frac{(1+\beta^2)\cdot \text{coherence}\cdot \text{min\_edit}}{\beta^2 \cdot \text{coherence} + \text{min\_edit}}$, as a function of the scaling factor β. Low β values assign more weight to coherence, and high β values emphasize more on min-edit.

5 Task 2: Counterfactual Reasoning

Given an original story ending Z of story context X^{ori}, and a counterfactual condition X that changes X^{ori} to invalidate Z (see Fig. 1), the task is to generate a new story ending Y that minimally edits the original ending Z to regain coherence with the counterfactual condition X (Qin et al., 2019a).

5.1 Task Setup

Constraints The constraint we enforce is that Y is close to Z (i.e., minimal edits). We impose this constraint by minimizing their KL divergence:

$$\mathcal{L}(X, \tilde{Y}, Z) := \text{KL}\left(Z \| \text{softmax}(\tilde{Y}/\tau)\right), \quad (7)$$

where, with a slight abuse of notation, Z is the one-hot distribution of the tokens in the original ending. That is, we encourage the generated logits to recover the original ending.

Ranking We rank the candidates based on both their coherence with the context, as well as the internal coherence between the multiple sentences of each candidate (rewritten ending, consists of 3 sentences). More concretely, given a candidate Y, we compute the aggregated coherence score:

$$\text{ranking\_score}(Y) = c(X, Y) + \sum_{s=1}^{S-1} c(Y[s], Y[s+1]), \quad (8)$$

where each candidate has S sentences (here, $S = 3$) and $Y[s]$ denotes the sth sentence.
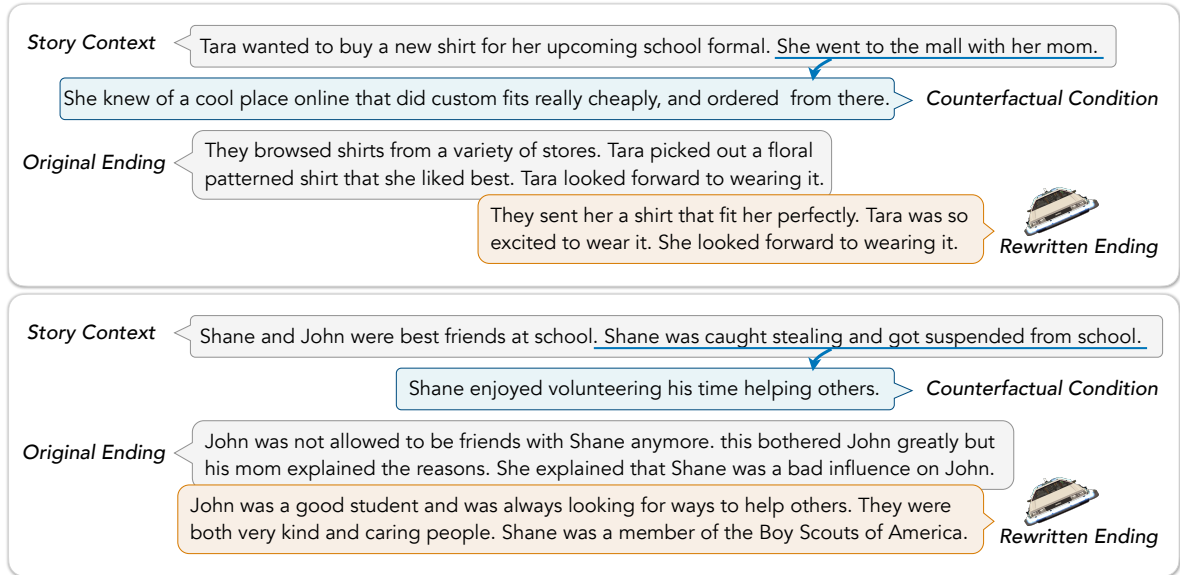
Figure 5: Examples of generated story endings on three counterfactual reasoning cases. Given a story context, a counterfactual condition, and a original ending, DELOREAN generates a rewritten ending which is coherent with the counterfactual condition and is similar to the original ending.

Hyperparameters We largely follow the same setting as in the abductive reasoning task, but tune hyperparameters on the TIMETRAVEL development set. Deviations from these settings are outlined in Appendix A.2.

5.2 Experimental Setup

Baselines We compare our method with baselines from Qin et al. (2019a). The zero-shot baseline uses the pre-trained GPT-2 model to generate Y as a continuation to the counterfactual condition X. It is the most apt comparison to our method which also doesn't require additional supervision. We also experiment with two baselines that fine-tune GPT-2 on the original story $X^{ori}Z$ to fit the model to the story domain, either with an LM objective (FT) or a tailored conditional objective that encourages minimal edits of Z (Recon+CF).[6] Finally, we report the performance of a supervised baseline (Sup), in which GPT-2 is fine-tuned to produce the gold Y from $X^{ori}Z$ and X.

5.3 Results

Automatic Evaluation Following Qin et al. (2019a), we report BERTSCORE (Zhang et al., 2019), which was shown to best correlate with human judges' notion of counterfactual coherence, and BLEU-4 and ROUGE-L, which better measure minimum-edits. We find that the discriminative baselines achieve the highest degree of plot

fidelity. Meanwhile, DELOREAN achieves the highest BERTSCORE for counterfactual coherence.

Human Evaluation We repeat the human evaluation setup from Section 4.3. Presented with the original story, the counterfactual condition X, and the generated ending Y, workers were asked to judge (1) the *coherence* of Y with respect to the X; and (2) to what extent the generated ending *minimally-edits* the original ending.[7] In order to judge both criteria, we report the weighted harmonic mean H_β of these scores across a range of weights β (Figure 4).

Our results show that DELOREAN is the only model that maintains a consistent balance between *coherence* (1.66) and *minimal edits* (1.54). While the ranking-augmented zero-shot model produces the most coherent endings (*coherence* = 1.8), it deviates from the original ending. As β is increased (i.e., increasing importance of *minimal edits*), its weighted performance drops considerably, indicating it cannot generate new endings that follow the original plot of the story (*min-edit* = 1.25). Conversely, Recon+CF generates stories that are faithful to the original endings, but are far less coherent with the counterfactual condition (*coherence* = 1.23). Through human annotation, we found that Recon+CF copies the original ending word-for-word in a 84% of cases.

The pairwise comparison results in Table 5

---

[6]See Qin et al. (2019a) for more details.

[7]Fair inter-rater agreement with Fleiss' $\kappa = 0.34$

parallel these observations. DELOREAN significantly outperforms the discriminative approaches (Recon+CF and Sup+Disc) in coherence, while falling short of the Zero-shot re-ranked baselines. In minimal edits, this pattern is flipped with our approach outperforming Zero-shot baselines considerably and losing to the discriminative baselines.

Qualitative Analysis Figure 5 provides two example results for counterfactual story rewriting by DELOREAN. The approach successfully captures the causal relations between events and properly rewrites the endings with minimal edits. For instance, in the first example, given the counterfactual condition that "Tara ordered a shirt online" (as opposed to the original "went to mall"), the rewritten ending is about "sent shirt" to Tara (as opposed to the original "browsed from stores"). The last sentence of the original ending "She looked forward to wearing it" is correctly preserved as it is coherent with the counterfactual condition.

6 Related Work

Unsupervised text generation. Unsupervised approaches are often applied to problems that copy information from a source text into decoded text. Unsupervised paraphrasing requires repeating this information (Miao et al., 2019; Bao et al., 2019), as does translation, but with a bilingual transformation (Artetxe et al., 2017; Lample et al., 2018). In summarization there is an additional task to select a subset of the original text (Baziotis et al., 2019; Schumann et al., 2020; West et al., 2019). In cases where information is mostly copied from the original, auto-encoding objectives can ensure the correct information is captured (Bao et al., 2019; Baziotis et al., 2019; Artetxe et al., 2017). This work tackles problems where generation is more open-ended. Rather than reproducing information from the prompt, generations should agree with and expand on it, making autoencoding less applicable.

Controllable language generation. Earlier approaches for controllable generation involved preserving the content of text while changing it along discrete dimensions, such as theme, sentiment, or style (Koncel-Kedziorski et al., 2016; Hu et al., 2017; Ficler and Goldberg, 2017; Shen et al., 2017; Lample et al., 2019). Recent works such as Grover (Zellers et al., 2019) and CTRL model (Keskar et al., 2019) used these ideas to augment transformer language models that can condition on structured metadata such as source, domain, etc. The Plug & Play model (PPLM; Dathathri et al., 2019) controls topic and sentiment in an approach similar to ours that involves forward and backward passes to update token distributions. However, PPLM relies on trained attribute discriminators for supervision, while our method is unsupervised. While these models are restricted to specific dimensions, often with pre-defined values, our model can adjust to any open-ended textual constraint. Perhaps the most similar work in that aspect is the "text infilling" models, which, however, are in a more narrow setting by filling only a relatively short text span (Devlin et al., 2018; Zhu et al., 2019; Donahue et al., 2020), and more restrictive due to the reliance on an extra right-to-left language model (Sun et al., 2017) or a pre-specified generation length (Zeldes et al., 2020, which is not publicly available).

Reasoning about narratives. A prominent resource from recent years is the RocStories corpus (Mostafazadeh et al., 2016b), consisting of 98K crowdsourced 5-sentence everyday life stories. It was used for the story cloze task whose goal was to predict the story ending from its first 4 sentences, but gained popularity and became the base of additional benchmarks (Rashkin et al., 2018). Additional related work includes "script knowledge", i.e. learning about prototypical series of events (Schank and Abelson, 1977; Chambers and Jurafsky, 2008; Pichotta and Mooney, 2014), temporal commonsense (Granroth-Wilding and Clark, 2016; Li et al., 2018), and modeling pre- and post- conditions of events (Roemmele et al., 2011; Sap et al., 2019; Bosselut et al., 2019). Qin et al. (2019b) studied conversation modeling that reads and connects the dots of events in related documents. Finally, a recent line of work explores counterfactual questions in reading comprehension (Huang et al., 2019; Tandon et al., 2019), but instantiates the problem of counterfactual reasoning as a multiple choice task.

7 Conclusion

We presented DELOREAN, an unsupervised LM-based approach to generate text conditioned on past context as well as future constraints, through forward and backward passes considering each condition. We demonstrated its effectiveness for abductive and counterfactual reasoning, on which it performed substantially better than unsupervised baselines. Our method is general and can be easily adapted for other generative reasoning tasks.

References

Henning Andersen. 1973. Abductive and deductive change. *Language*, pages 765–793.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. *arXiv preprint arXiv:1907.05789*.

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. Seq3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *NAACL-HLT*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

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.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

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

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: a simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

C. Donahue, M. Lee, and P. Liang. 2020. Enabling language models to fill in the blanks. In *ACL*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *ACL*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104.

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

Matthew L Ginsberg. 1986. Counterfactuals. *Artificial intelligence*, 30(1):35–79.

Nelson Goodman. 1947. The problem of counterfactual conditionals. *The Journal of Philosophy*, 44(5):113–128.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Jerry R Hobbs, Mark E Stickel, Douglas E Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial intelligence*, 63(1-2):69–142.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.

Steve D Isard. 1974. What would you have done if...? *Theoretical Linguistics*, 1(1-3):233–256.

Philip Nicholas Johnson-Laird. 2006. *How we reason*. Oxford University Press, USA.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Rik Koncel-Kedziorski, Ioannis Konstas, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2016. A theme-rewriting approach for generating algebra word problems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1617–1628.

Sarit Kraus, Daniel Lehmann, and Menachem Magidor. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial intelligence*, 44(1-2):167–207.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *ICLR*.

Carolin Lawrence and Stefan Riezler. 2018. Improving a neural semantic parser by counterfactual learning from human bandit feedback. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1820–1830.

Carolin Lawrence, Artem Sokolov, and Stefan Riezler. 2017. Counterfactual learning from bandit feedback under deterministic logging: A case study in statistical machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2566–2576.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *IJCAI*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Hugo Mercier and Dan Sperber. 2017. *The enigma of reason*. Harvard University Press.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6834–6842.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016b. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *HLT-NAACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books.

Charles Sanders Peirce. 1960. *Collected papers of charles sanders peirce*, volume 2. Harvard University Press.

Karl Pichotta and Raymond Mooney. 2014. Statistical script learning with multi-argument events. In *EACL*, pages 220–229.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019a. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5046–5056.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019b. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *ACL*, pages 5427–5436.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners. -.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019b. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*.

Raymond Reiter. 1988. Nonmonotonic reasoning. In *Exploring artificial intelligence*, pages 439–481. Elsevier.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

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. 1977. Scripts, plans, goals and understanding: An inquiry into human knowledge structures.

Raphael Schumann, Lili Mou, Yao Lu, Olga Vechtomova, and Katja Markert. 2020. Discrete optimization for unsupervised sentence summarization with word-level extraction. *arXiv preprint arXiv:2005.01791*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Youngseo Son, Anneke Buffone, Joe Raso, Allegra Larche, Anthony Janocko, Kevin Zembroski, H Andrew Schwartz, and Lyle Ungar. 2017. Recognizing counterfactual thinking in social media texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 654–658.

William Starr. 2019. Counterfactuals. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2019 edition. Metaphysics Research Lab, Stanford University.

Qing Sun, Stefan Lee, and Dhruv Batra. 2017. Bidirectional beam search: Forward-backward inference in neural sequence models for fill-in-the-blank image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6961–6969.

Niket Tandon, Bhavana Dalvi Mishra, Keisuke Sakaguchi, Antoine Bosselut, and Peter Clark. 2019. Wiqa: A dataset for" what if..." reasoning over procedural text. In *EMNLP*.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi.

2019. Bottlesum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3743–3752.

Yoel Zeldes, Dan Padnos, and Barak Peleg. 2020. Haim-1.5 - the next generation.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Wanrong Zhu, Zhiting Hu, and Eric Xing. 2019. Text infilling. *arXiv preprint arXiv:1901.00158*.