Generating Hypothetical Events for Abductive Inference

Debjit Paul
Research Training Group AIPHES
Institute for Computational Linguistics
Heidelberg University
paul@cl.uni-heidelberg.de

Anette Frank
Research Training Group AIPHES
Institute for Computational Linguistics
Heidelberg University
frank@cl.uni-heidelberg.de

Abstract

Abductive reasoning starts from some observations and aims at finding the most plausible explanation for these observations. To perform abduction, humans often make use of temporal and causal inferences, and knowledge about how some hypothetical situation can result in different outcomes. This work offers the first study of how such knowledge impacts the *Abductive* α***NLI*** task – which consists in choosing the more likely explanation for given observations. We train a specialized language model $LM_{\mathcal{I}}$ that is tasked to generate *what could happen next* from a hypothetical scenario that evolves from a given event. We then propose a multi-task model $\mathcal{MTL}$ to solve the α**NLI task**, which predicts a plausible explanation by a) considering different *possible events* emerging from candidate hypotheses – events generated by $LM_{\mathcal{I}}$ – and b) selecting the one that is most *similar* to the observed outcome. We show that our $\mathcal{MTL}$ model improves over prior vanilla pre-trained LMs fine-tuned on αNLI. Our manual evaluation and analysis suggest that learning about possible next events from different hypothetical scenarios supports abductive inference.

1 Introduction

Abductive reasoning (AR) is inference to the best explanation. It typically starts from an incomplete set of observations about everyday situations and comes up with what can be considered the most likely possible explanation given these observations (Pople, 1973; Douven, 2017). One of the key characteristics that make abductive reasoning more challenging and distinct from other types of reasoning is its non-monotonic character (Strasser and Antonelli, 2019) i.e., even the most likely explanations are not necessarily correct. For example, in Figure 1, the most likely explanation for *Observation 1: "wet grass outside my house"* is that *"it has been raining"*. However, when a new piece of information (observation or evidence) becomes available, the explanation must possibly be retracted, *showing the defeasible character of abduction*. With the new observation (*"the sprinkler was switched on"*) the most plausible explanation changes to *"Sprinkler caused the grass to be wet"*. Humans, in such situations, could induce or validate such abductive inferences by performing hypothetical reasoning (such as *"What would happen if the sprinkler was switched on?"*) to arrive at a plausible explanation for *"wet grass outside my house"*.

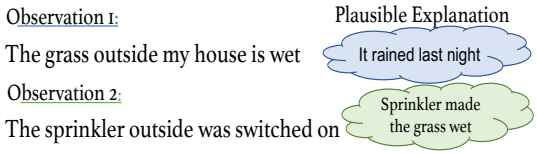

Figure 1: Motivational example illustrating Abductive Reasoning and its non-monotonic character.

In this work, we focus on the α**NLI task** (Bhagavatula et al., 2020), where given two observations (O_1 at time t_1, O_2 at time t_2, with $t_1 < t_2$) as an incomplete context, the task is to predict which of two given hypothesized events (H_1 or H_2) is more plausible to have happened between O_1 and O_2. Figure 2 illustrates this with an example: given observations O_1*:"Priya decided to try a new restaurant."* and O_2*: "Priya thought her food was delicious."*, the task is to predict whether H_1 or H_2 is the more plausible explanation given observations O_1 and O_2. Both H_1 and H_2 are different plausible hypothetical situations that can evolve from the same observation (premise) O_1.

In this paper, we hypothesize that learning how different hypothetical scenarios (H_1 and H_2) can result in different outcomes (e.g., $O_2^{H_j}$, Fig. 2) can help in performing abductive inference. In order to decide which H_i, is *more plausible* given observa-

Proceedings of the 10th Conference on Lexical and Computational Semantics, pages 67–77
August 5–6, 2021, Bangkok, Thailand (online)

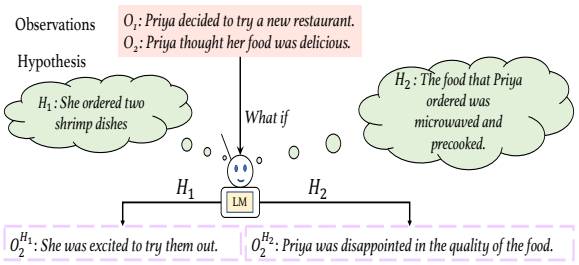

Figure 2: Motivational example for αNLI : The top box (red) shows the observations and two callout clouds (green) contain the hypotheses. The implications ($O_i^{H_i}$) – generated by the LM conditioned on each hypothesis and the observations – are given in pink colored boxes.

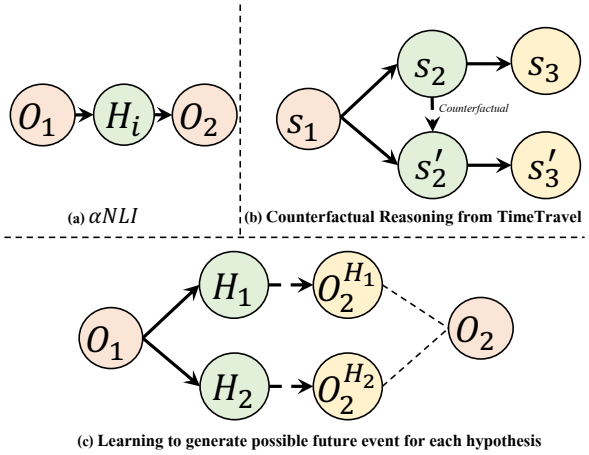

Figure 3: Different reasoning schemes and settings for our task and approach. The arrows denote the direction (temporal flow) of the reasoning chain. The dotted arrow in (b) denotes the derivation of a counterfactual situation s_2' from a factual s_2. In (c), the dotted arrows denote the learned inference; the dotted lines indicate the similarity between O_2 and $O_2^{H_i}$.

tions, we assume each H_i to be *true* and generate a *possible next event* $O_2^{H_i}$ for each of them independently (e.g.: *What will happen if Priya's ordered food was microwaved and precooked?*). We then compare the generated sentences ($O_2^{H_1}$, $O_2^{H_2}$ in Fig. 2) to what has been observed (O_2) and choose as most plausible hypothesis the one whose implication is closest to observation O_2.

We design a language model ($LM_{\mathcal{I}}$) which, given observations and a hypothesis, generates a possible event that could happen next, given one hypothesis. In order to train this language model, we use the TIMETRAVEL (TT) corpus (Qin et al., 2019) (a subpart of the *ROCStories* corpus[1]). We utilize the $LM_{\mathcal{I}}$ model to generate a possible next event for each hypothesis, given the observations. We then propose a multi-task learning model $\mathcal{MTL}$ that jointly chooses from the generated possible next events ($O_2^{H_1}$ or $O_2^{H_2}$) the one most similar to the observation O_2 and predicts the most plausible hypothesis (H_1 or H_2).

Our contributions are: i) To our best knowledge, we are the first to demonstrate that a model that learns to perform hypothetical reasoning can support and improve abductive tasks such as αNLI. We show that ii) for αNLI our multi-task model outperforms a strong BERT baseline (Bhagavatula et al., 2020).

Our code is made publicly available.[2]

2 Learning about Counterfactual Scenarios

The main idea is to learn to generate assumptions, in a given situation, about *"What could have happened (next) if we had done X?" or "What could happen (next) if we do X?"* (Bhatt and Flanagan, 2010). Figure 3(a) depicts the αNLI task framework. We hypothesize that getting to know *what will happen (next) if any of two hypotheses occurs*, will help us verifying which of them is more plausible (see Fig. 3(c)). Therefore, we encourage the model to learn how different hypothetical events (including counterfactual events) evolving from the same premise (s_1) can lead to different or similar outcomes (see Fig. 3(b)).

Accordingly, we teach a pre-trained GPT-2 (Radford et al., 2019) language model how to generate *a sequence of* possible subsequent events given different hypothetical situations in a narrative setting. Training such a model on narrative texts encourages it to learn causal and temporal relations between events. We train a conditional language model, $LM_{\mathcal{I}}$, which generates a possible event that could happen next, given some counterfactual scenarios for a given story.

We train this model on the TIMETRAVEL (TT) dataset (Qin et al., 2019), by fine-tuning GPT-2 to learn about possible next events emerging from a situation in a story, given some alternative, counterfactual event. The TT dataset consists of five-sentence instances S=(s_1,s_2,..,s_5)[3] from the ROCStories corpus[1] plus additional crowd-sourced sen-

[1] We ensure that αNLI testing instances are held out.

[2] https://github.com/Heidelberg-NLP/HYPEVENTS

[3] s_1 = *premise*, s_2 = *initial context*, $s_{3:5}$ = original ending

O_1: *Dotty was being very grumpy.*
O_2: *She felt much better afterwards.*
H_1: Dotty ate something bad.
H_2: **Dotty call some close friends to chat.**
$O_2^{H_1}$: *She started to feel sick.*
$O_2^{H_2}$: *They all tried to make her happy.*

Table 1: Example of generated possible next events $O_2^{H_j}$ using the $LM_{\mathcal{I}}$ model. **Bold** hypothesis (H_2) is more plausible.

tences $s^{'}_{2:5}$, where $s^{'}_2$ is counterfactual[4] to s_2 from the original story[5]. There are two reasons for using the TT dataset for our purposes: a) the domains on which GPT-2 was pretrained are broad[6] and different from the domain of ROCStories, b) the model can see how alternative situations can occur starting from the same premise s_1, resulting in similar or different outcomes. Note that, although intermediate situations may be counterfactual to each other, the future outcome can still be similar to the original ending due to *causal invariance* [7].

Concretely, the language model $LM_{\mathcal{I}}$ reads the premise (s_1) and the alternative event(s) (s_2 or $s^{'}_2$), the masked token (serving as a placeholder for the missing possible next event(s) ($s_{3:i}$ or $s^{'}_{3:i}$), then the rest of the story ($s_{i+1:5}$ or $s^{'}_{i+1:5}$) and again the premise (s_1). We train the model to maximize the log-likelihood of the missing ground-truth sentence(s) ($s_{3:i}$).

$$\mathcal{L}^{LM_{\mathcal{I}}}(\beta) = log_{p_\beta}(s_{3:i}|[S]s_1,[M],s_{i+1:5},[E],[S],s_1,s_2) + log_{p_\beta}(s^{'}_{3:i}|[S]s_1,[M],s^{'}_{i+1:5},[E],[S],s_1,s^{'}_2) \quad (1)$$

where $i \in [3,4]$, s_i=$\{w_1^{s_i},..,w_n^{s_i}\}$ a sequence of tokens, $[S]$=start-of-sentence token, $[E]$=end-of-sentence token, $[M]$=mask token.

3 Generating Hypothetical Events to support the αNLI task

In this paper, we aim to investigate whether models perform better on the αNLI task when explicitly learning about events that could follow other events in a hypothetical scenario. We do so by introducing two methods $LM_{\mathcal{I}}$ + *BERTScore* and $LM_{\mathcal{I}}$ + $\mathcal{MTL}$ for unsupervised and supervised settings, respectively.

We first apply the trained model $LM_{\mathcal{I}}$ on the αNLI task, where the given observations O_1 and O_2, and alternative hypotheses H_j are fed as shown in (2) below.[8]

$$O_2^{H_j} = \beta([S],O_1,[M],O_2,[E],[S],O_1,H_j) \quad (2)$$

We generate a possible next event for each hypothetical event H_j, i.e., $O_2^{H_1}$ and $O_2^{H_2}$ (or: what will happen if some hypothesis H_j occurs given the observations), where $j \in [1,2]$. Table 1 illustrates an example where different $O_2^{H_j}$ are generated using $LM_{\mathcal{I}}$. One of the challenges when generating subsequent events given a hypothetical situation is that there can be infinite numbers of possible next events. Therefore, to constrain this range, we chose to give future events (O_2) as input, such that the model can generate subsequent events in a constrained context.

3.1 Unsupervised Setting

In this setting, we do not train any supervised model to explicitly predict which hypothesis is more plausible given the observations. Instead, we apply the fine-tuned $LM_{\mathcal{I}}$ model to the αNLI data, generate possible next events $O_2^{H_j}$ given O_1 and H_j, as described above, and measure the similarity between such possible next events ($O_2^{H_j}$) and the observation (O_2) in an unsupervised way, using *BERTScore* (BS) (Zhang et al., 2020) [9]. We evaluate our hypothesis that the generated possible next event $O_2^{H_j}$ given the more plausible hypothesis H_j should be *more similar* to observation O_2. Table 1 illustrates an example where H_2 is the more plausible hypothesis. We impose the constraint that for a correctly predicted instance BS($O_2{}^{H^+}$, O_2) $>$ BS($O_2{}^{H^-}$, O_2) should hold, where H^+, H^- are the more plausible vs. implausible hypothesis, respectively.

3.2 Supervised Setting

In this setting, displayed in Figure 4, we explore the benefits of training a multi-task $\mathcal{MTL}$ model that predicts i) the most plausible hypothesis and ii) which possible next event ($O_2^{H_j}$) is more similar

[4] a counterfactual $s^{'}$ states something that is contrary to s

[5] During our experiments we treat them as two separate instances: S_1=($s_{1:5}$) and S_2 = (s_1,$s^{'}_{2:5}$).

[6] GPT-2 was trained on the WebText Corpus.

[7] the future events that are invariant under the counterfactual conditions (Qin et al., 2019)

[8] For definition of placeholders see (1).

[9] BERTScore is an automatic evaluation metric for text generation that leverages the pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.

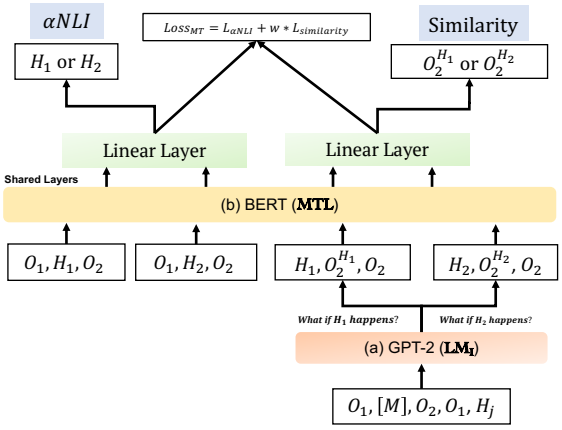

Figure 4: Overview of our $LM_{\mathcal{I}}$ + $\mathcal{MTL}$ model for αNLI: (a) language model $LM_{\mathcal{I}}$ takes the input in a particular format to generate different possible next events, (b) the $\mathcal{MTL}$ model learns to predict the best explanation (H_j) and possible next events ($O_2^{H_j}$) at the same time to perform the αNLI task.

to the observation (O_2). Multi-task learning aims to improve the performance of a model for a task by utilizing the knowledge acquired by learning related tasks (Ruder, 2019). We *hypothesize that* a) the possible next event $O_2^{H_j}$ of the more plausible hypothesis H_j should be most similar to observation O_2, and that b) learning which possible next event is more similar supports the model in the αNLI task (*inductive transfer*).

The architecture of $LM_{\mathcal{I}}$ + $\mathcal{MTL}$ model is shown in Figure 4. The model marked (a) in Figure 4 depicts the $LM_{\mathcal{I}}$ model as described in §3. The outputs of the $LM_{\mathcal{I}}$ model, which we get from Eq. (2) for both hypotheses are incorporated as an input to the $\mathcal{MTL}$ model. Concretely, we feed the $\mathcal{MTL}$ classifier a sequence of tokens as stated in part (b) of Figure 4, and aim to compute their contextualized representations using pre-trained BERT. The input format is described in Table 3. Similar to (Devlin et al., 2019), two additional tokens are added [CLS] at the start of each sequence input and [SEP] at the end of each sentence. In the shared layers (see Fig 4(b)), the model first transform the input sequence to a sequence of embedding vectors. Then it applies an attention mechanism that learns contextual relations between words (or sub-words) in the input sequence.

For each instance we get four [CLS] embeddings ($CLS_{H_j}, CLS_{O_2^{H_j}}$; $j \in [1, 2]$) which are then passed through two linear layers, one for the αNLI (main task) and another for predicting the similarity (auxiliary task) between $O_2^{H_j}$ and O_2. We compute the joint loss function $\mathcal{L}$ = $\mathcal{L}_{\alpha NLI}$ + $w * \mathcal{L}_{similarity}$; where w is a trainable parameter, $\mathcal{L}_{\alpha NLI}$ and $\mathcal{L}_{similarity}$ are the loss function for the αNLI task and auxiliary task, respectively.

Task	Train	Dev	Test
αNLI	169654	1532	3059
TimeTravel (NLG)	53806	2998	–

Table 2: Dataset Statistics: nb. of instances

Input Format	Output
[CLS] O_1 [SEP] H_i [SEP] O_2 [SEP]	H_1 or H_2
[CLS] H_i [SEP] $O_2^{H_i}$ [SEP] O_2[SEP]	$O_2^{H_1}$ or $O_2^{H_2}$

Table 3: Input and output format for the αNLI task: [CLS] is a special token used for classification, [SEP] a delimiter.

4 Experimental Setup

Data. We conduct experiments on the $\mathcal{ART}$ (Bhagavatula et al., 2020) dataset. Data statistics are given in Table 2. For evaluation, we measure accuracy for αNLI.

Hyperparameters. To train the $LM_{\mathcal{I}}$ model we use learning rate of $5e-05$. We decay the learning rate linearly until the end of training; batch size: 12. In the supervised setting for the α**NLI task**, we use the following set of hyperparameters for our $\mathcal{MTL}$ model with integrated $LM_{\mathcal{I}}$ model ($LM_{\mathcal{I}}$ + $\mathcal{MTL}$): batch size: $\{8, 16\}$; epochs: $\{3, 5\}$; learning rate: $\{2e\text{-}5, 5e\text{-}6\}$. For evaluation, we measure accuracy. We use Adam Optimizer, and dropout rate = 0.1. We experimented on GPU size of 11GB and 24GB. Training is performed using cross-entropy loss. The loss function is $\mathcal{L}_{\alpha NLI}$ + $w * \mathcal{L}_{similarity}$, where w is a trainable parameter. During our experiment we initialize $w = 1$. The input format is depicted in Table 3. We report performance by averaging results along with the variance obtained for 5 different seeds.

Baselines. We compare to the following baseline models that we apply to the αNLI task, training them on the training portion of the $\mathcal{ART}$ dataset (cf. Table 2).

- *ESIM + ELMo* is based on the ESIM model previously used for NLI (Chen et al., 2017). We use (a) ELMo to encode the observations and hypothesis, followed by (b) an attention

Model	Dev Acc.(%)	Test Acc.(%)
Majority (*from dev set*)$^{\diamond}$	–	50.8
$LM_{\mathcal{I}}$ + BERTScore	62.27	**60.08**
Infersent $^{\diamond}$	50.9	50.8
ESIM + ELMo $^{\diamond}$	58.2	58.8
BERT$_{Large}$ $^{\diamond}$	69.1	68.9±0.5
GPT-2 +$\mathcal{MTL}$	68.9±0.3	68.8±0.3
COMET +$\mathcal{MTL}$	69.4±0.4	69.1±0.5
$LM_{\mathcal{I}} + \mathcal{MTL}$	72.9±0.5	**72.2**±0.6
Human Performance	-	91.4

Table 4: Results on α**NLI task**, $\diamond$: as in Bhagavatula et al. (2020) (no unpublished leaderboard results). For each row, the best results are in bold, and performance of our models are in blue.

layer, (c) a local inference layer, and (d) another bi-directional LSTM inference composition layer, and (e) a pooling operation,

- *Infersent* (Conneau et al., 2017) uses sentence encoding based on a bi-directional LSTM architecture with max pooling.
- *BERT* (Devlin et al., 2019) is a LM trained with a masked-language modeling (MLM) and next sentence prediction objective.

As baselines for using the $\mathcal{MTL}$ model, we replace $LM_{\mathcal{I}}$ with alternative generative LMs:

- *GPT-2* + $\mathcal{MTL}$. In this setup, we directly use the pretrained GPT-2 model and task it to generate a next sentence conditioned on each hypothesis ($O_2^{H_i}$) without finetuning it on the TIMETRAVEL data. We then use the supervised $\mathcal{MTL}$ model to predict the most plausible hypothesis and which of the generated observations is more similar to O_2.
- *COMET* + $\mathcal{MTL}$. In this setting, we make use of inferential *if-then* knowledge from ATOMIC (Sap et al., 2019a) as background knowledge. Specifically, we use COMET to generate objects with **Effect**[10] relations for each hypothesis as a textual phrase.

5 Results

In Table 4, we compare our models $LM_{\mathcal{I}}$ + BERTScore and $LM_{\mathcal{I}} + \mathcal{MTL}$ against the models proposed in Bhagavatula et al. (2020): a majority baseline, supervised models (*Infersent* and *ESIM+ELMo*), as well as *BERT*$_{Large}$. Bhagavatula et al. (2020) re-train the ESIM+ELMo and Infersent models on the $\mathcal{ART}$ dataset and fine-tuned the BERT model on the αNLI task and report the results.

We find that our **unsupervised** model with BERTScore ($LM_{\mathcal{I}}$ + BERTScore) outperforms (by +9.28 pp. and +1.28 pp.) strong ESIM+ELMo and Infersent baseline models. Table 5 shows some examples of our generation model $LM_{\mathcal{I}}$ along with the obtained BERTScores.

Unlike the unsupervised $LM_{\mathcal{I}}$ + BERTScore, our **supervised** $LM_{\mathcal{I}} + \mathcal{MTL}$ model also improves over the BERT$_{Large}$ baseline, by +3.3 pp. We can attribute the improvement to the model having been jointly trained to assess the similarity and dissimilarity of possible next events $O_2^{H_j}$ and observations (O_2) along with the αNLI task. One of the advantages of training our proposed multi-task learning ($\mathcal{MTL}$) model, instead of directly feeding the possible next events $O_2^{H_j}$ as knowledge inputs is that it adds an explainable component to the model. One can view the generated next events $O_2^{H_j}$ as natural language rationales and our multi-task model explicitly chooses one of them. Hence, the multi-task framework makes the model more expressive. Finally, we compare, for the $\mathcal{MTL}$ model, our embedded generation model $LM_{\mathcal{I}}$ to pre-trained GPT-2 and COMET. Table 4 shows that $LM_{\mathcal{I}}$ + $\mathcal{MTL}$ yields better performance compared to both *COMET* + $\mathcal{MTL}$ (+3.1 pp.) and *GPT-2* + $\mathcal{MTL}$ (+3.4 pp.) – the intuitive reason being that the next events generated by $LM_{\mathcal{I}}$ are more helpful than events generated using pretrained GPT-2 and objects generated by COMET.

Table 5 illustrates some examples where our $\mathcal{MTL}$ model not only chooses the correct hypothesis, but also a likely possible next event that is similar to the observation O_2. Interestingly, during training of $\mathcal{MTL}$ we initialize w = 1, and after training the model we found the w value had been adjusted to a range between 0.85 and 0.75, which intuitively shows both the effectiveness of our $LM_{\mathcal{I}}$ -generated possible next events, and their similarity to the given observations O_2.

5.1 Case Study

Table 5 displays possible next events, generated by our $LM_{\mathcal{I}}$ model – along with the BERTscore measured between the possible next events $O_2^{H_j}$ and observation O_2. We see two different scenarios:

[10] as a result PersonX feels; as a result PersonX wants; PersonX then

	Observation (O)	Given Hypotheses (H_j)	What *if* H_j happens? Generated possible next events $O_2^{H_j}$	MTL	Reasoning Type	BERT Score	Human Evaluation
a.	O_1: Jim got ready for his first date.	H_1: **Jim's date wasn't attracted to him.**	$O_1^{H_2}$: He tried to get her number. She never responded to him.	$H_1, O_2^{H_1}$	*Reaction*	**47.34**	Non-Contradiction
	O_2: Since then, she has ignored all of Jim's text messages.	H_2: Jim went on the date and said he didn't like the girl.	$O_2^{H_2}$: He told her that he didn't want to date her.			44.19	Contradiction
b.	O_1: Nate had the summer off before college.	H_1: **Nate spent the summer traveling and partying.**	$O_2^{H_1}$: Nate had a lot of fun.	$H_1, O_2^{H_1}$	*Motivation*	**49.02**	Non-Contradiction
	O_2: Nate's last summer before college was a total blast!	H_2:Nate decided to spend the entire summer working in the Mines.	$O_2^{H_2}$: Nate spent the summer working in the mines.			44.34	Redundancy
c.	O_1: Dotty was being very grumpy.	H_1: Dotty ate something bad.	$O_2^{H_1}$: She started to feel sick.	$H_2, O_2^{H_2}$	*Emotional Effect*	**0.604**	Contradiction
	O_2: She felt much better afterwards	H_2: **Dotty call some close friends to chat.**	$O_2^{H_2}$: They all tried to make her happy.			0.405	Non-Contradiction
d.	O_1: Cay had a crush on a boy in her class.	H_1: **Cay sent a love note to the boy.**	$O_2^{H_1}$: The boy responded and said he liked Cay.	$H_1, O_2^{H_1}$	*Emotional Effect*	**0.509**	Non-Contradiction
	O_2: He smiled at her after and said he liked her too!	H_2: She told him she did not like him.	$O_2^{H_2}$: The boy was very sad about it.			0.423	Contradiction

Table 5: Examples of generated possible next events for solving αNLI using our $LM_{\mathcal{I}}$ model. Column 3: Hypothesis and possible next events chosen by our $LM_{\mathcal{I}} + \mathcal{MTL}$ model; Column 4: Reasoning type between the hypothesis H_j and O_2; Column 5: BERTScore between the $O_2^{H_j}$ and O_2; Column5: Human evaluation of the possible next events with respect the observation O_2.

(i) examples (a), (b) and (d) depicting the scenario where possible next events and observation pairs *correctly* achieve higher BERTscores [11], (ii) example (c) depicting the scenario where an incorrect possible next event and observation pair achieves higher BERTscores than the correct one. Intuitive reasons for these scenarios are, for example, for (a): there is a higher word overlap and semantic similarity between a correct next event and observation O_2, for example (b): there is higher semantic similarity; whereas for example (c): although there is a higher semantic dissimilarity, the word overlap between the wrong possible next event (*"She started to feel sick."*) and the observation (*"She felt much better afterwards."*) is much higher.

6 Manual Evaluation

Since the automatic scores only account for word-level similarity between observations and generated possible next events, we conduct a manual evaluation study, to assess the quality of sentences generated by our $LM_{\mathcal{I}}$ model.

Annotation Study on $LM_{\mathcal{I}}$ generations. The annotation was performed by three annotators with computational linguistic background. We provide each of the three annotators with observations, hypotheses and sentences, as produced by our $LM_{\mathcal{I}}$ model, for 50 randomly chosen instances from the αNLI task. They obtain i) *generated sentences for a next possible event* for the *correct* and *incorrect hypothesis*, as well as ii) the *sentence stating observation* O_2.

We ask each annotator to rate the sentences according to four quality aspects as stated below.

Grammaticality: the sentence is i) grammatical, ii) not entirely grammatical but understandable, or iii) completely not understandable;

Redundancy: the sentence contains redundant or repeated information;

Contradiction: the sentence contains any pieces of information that are contradicting the given observation O_2 or not;

Relevance: the possible next event is i) relevant, ii) partially relevant, or iii) not relevant.

For each aspect, they are asked to judge the sentence generated for the correct hypothesis[12]. Only for **Contradiction**, they are asked to judge both sentences, for correct and the incorrect hypotheses.

Results and Discussion. Figures 5, 7, and 6 present the results of manual evaluations of the generation quality, according to the different criteria described above.

[11] BERTscore matches words in candidate and reference sentences by cosine similarity.

[12] The correct hypothesis was marked for the annotation.

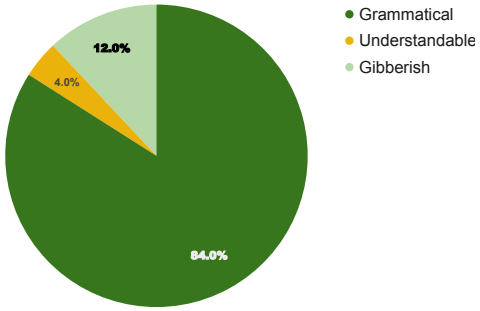

Figure 5: Human evaluation of the *grammaticality* of generated sentences: ratio of i) grammatical, ii) not entirely grammatical but understandable, iii) completely not understandable sentences.

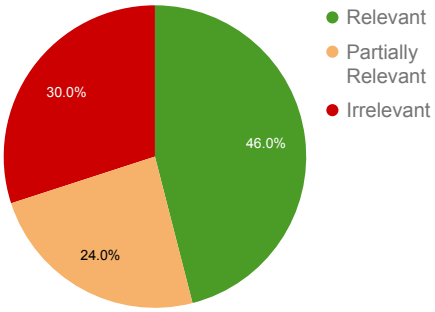

Figure 6: Human evaluation of the *Relevance* of generated sentences for possible next events.

For measuring inter-annotator agreement, we computed Krippendorff's α (Hayes and Krippendorff, 2007) for *Grammaticality* and *Relevance*, as it is suited for ordinal values, and Cohen's Kappa κ for *Redundancy* and *Contradiction*. We found α values are 0.587 and 0.462 for *Grammaticality* and *Relevance*, respectively (moderate agreement) and κ values 0.61 and 0.74 for *Redundancy* and *Contradiction* (substantial agreement). We aggregated the annotations from the three annotators using majority vote. Figure 5 shows that the majority of sentences (96%) are grammatical or understandable. Figure 7 shows that most sentences for correct labels are non-redundant (84%) and non-contradictory (88%), whereas for incorrect labels 39 instances are found to be contradictory with the observation O_2 (78%). The manual evaluation supports our hypothesis that the generated sentences for correct labels should be more similar (less contradictory) compared to the sentences generated for incorrect labels. Figure 6 shows the ratio of sentences considered by humans as relevant, partially relevant, and irrelevant. The results show that 46% of cases are relevant (based on majority agreement) and 24% of cases are partially relevant. This yields that the generated sentences are (partially) relevant

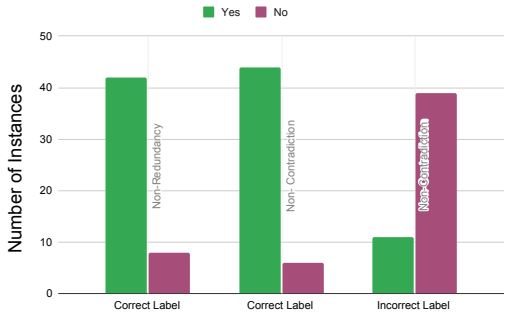

Figure 7: Human evaluation of *Redundancy* and *Contradiction* of generations for possible next events.

in most cases and thus should support abduction for both unsupervised ($LM_{\mathcal{I}}$ + BERTScore) and supervised ($LM_{\mathcal{I}}$ + $\mathcal{MTL}$) models.

Impact of Reasoning types. Finally, to better assess the performance of our model, we determine what *types of reasoning* underly the abductive reasoning tasks in our data, and examine to what extent our models capture or not these reasoning types. We consider again the 50 instances that were annotated by our previous annotators and manually classify them into different reasoning types. We broadly divided the data into 6 categories: (i) Motivation, (ii) Spatial-Temporal, (iii) Emotional, (iv) Negation, (v) Reaction, (vi) Situational fact. The most frequent type was Emotional (10), most infrequent was Spatial (7). We ask a new annotator to annotate the reasoning types for these 50 instances. Considering the relevance and contradiction categories from the previous annotations we determine that for Negation (8), Emotional (10), and Reaction (8) *all* generated events for *correct labels* are *partially or fully relevant and non-contradictory*. An intuitive reason can be that we train our $LM_{\mathcal{I}}$ model to learn how different counterfactual hypothetical events emerging from a single premise can lead to the same or different outcomes through a series of events. Some counterfactual events ($s_2^{'}$) are negations of the original event (s_2) in the TIMETRAVEL dataset. This may support the reasoning class Negation. For the other categories: Motivation, Spatial-temporal, and Situational fact, we detect errors regarding (missing) *Relevance* in 21%, 14% and 28% of cases, respectively. Table 6 illustrates an example from the class Situational Fact, where our generated next event is *irrelevant* and *redundant*.

O_1: *Jenna hit the weight hard in the gym.*
O_2: *She took a cold bath in order to alleviate her pain.*
H_1: Her neck pain stopped because of this.
H_2: **Jenna pulled a muscle lifting weights.**
$O_2^{H_1}$: *She decided to take a break .*
$O_2^{H_2}$: *Jenna lost weight in the gym.*

Table 6: Error Analysis: An example of generated possible next event $O_2^{H_j}$ from Situational Fact category.

7 Related Work

Commonsense Reasoning. There is growing interest in this research field, which led to the creation of several new resources on commonsense reasoning, in form of both *datasets*, such as SocialIQA (Sap et al., 2019b), CommonsenseQA (Talmor et al., 2019), CosmosQA (Huang et al., 2019) and *knowledge bases*, e.g. ConceptNet (Speer et al., 2017), ATOMIC (Sap et al., 2019a), or Event2Mind (Rashkin et al., 2018). Recently, many works proposed to utilize external *static* knowledge graphs (KGs) to address the bottleneck of obtaining relevant commonsense knowledge. Lin et al. (2019) proposed to utilize knowledge graph embeddings to rank and select relevant knowledge triples or paths. Paul and Frank (2019) proposed to extract subgraphs from KGs using graph-based ranking methods and further Paul et al. (2020) adopted the graph-based ranking method and proposed to *dynamically* extend the KG to combat sparsity. In concurrent work, Paul and Frank (2021) introduced a method to dynamically generate contextually relevant knowledge that guides a model while performing the narrative story completion task.

Both hypothetical reasoning and abductive reasoning are understudied problems in NLP. Recently, Tandon et al. (2019) proposed a first large-scale dataset of *"What if..."* questions over procedural text. They introduced the dataset to study the effect of perturbations in procedural text. Related to our work, Qin et al. (2019) investigated the capabilities of state-of-the-art LMs to rewrite stories with counterfactual reasoning. In our work we utilize this dataset to model how to generate possible next events emerging from different hypothetical and counterfactual events. Mostafazadeh et al. (2016) designed the narrative cloze task, a task to choose the correct ending of a story.[13] Conversely, Bhagavatula et al. (2020) proposed a task that requires reasoning about plausible explanations for narrative omissions. Our research touches on the issue of hypothetical reasoning about alternative situations. We found that making language models learn how different hypothetical events can evolve from a premise and result in similar or different future events forming from a premise and how these events can result in similar or different future events helps models to perform better in abduction.

Explainability. Despite the success of large pretrained language models, recent studies have raised some critical points such as: high accuracy scores do not necessarily reflect understanding (Min et al., 2019), large pretrained models may exploit superficial clues and annotation artifacts (Gururangan et al., 2018; Kavumba et al., 2019). Therefore, the ability of models to generate explanations has become desirable, as this enhances interpretability. Recently, there has been substantial effort to build datasets with natural language explanations (Camburu et al., 2018; Park et al., 2018; Thayaparan et al., 2020). There have also been numerous research works proposing models that are interpretable or explainable (Rajani et al., 2019; Atanasova et al., 2020; Latcinnik and Berant, 2020; Wiegreffe and Marasović, 2021). Our work sheds light in this direction, as our $\mathcal{MTL}$ model not only predicts the plausible hypothesis H_j but also generates possible next events $O_2^{H_j}$ and chooses the one that is closer to the given context, thereby making our model more expressive.

Abductive Reasoning. There has been longstanding work on theories of abductive reasoning (Peirce, 1903, 1965a,b; Kuipers, 1992, 2013). Researchers have applied various frameworks, some focused on pure logical frameworks (Pople, 1973; Kakas et al., 1992), some on probabilistic frameworks (Pearl, 1988), and others on Markov Logics (Singla and Mooney, 2011). Recently, moving away from logic-based abductive reasoning, Bhagavatula et al. (2020) proposed to study language-based abductive reasoning. They introduced two tasks: *Abductive Natural Language Inference (αNLI)* and *Generation (αNLG)*. They establish baseline performance based on state-of-the-art language models and make use of inferential structured knowledge from ATOMIC (Sap et al., 2019a) as background knowledge. Zhu et al. (2020) proposed to use a learning-to-rank framework to address the abductive reasoning task. Ji et al. (2020)

[13] Their dataset, ROCStories, was later extended in Qin et al. (2019) and Bhagavatula et al. (2020).

proposed a model GRF that enables pre-trained models (GPT-2) with dynamic multi-hop reasoning on multi-relational paths extracted from the external ConceptNet commonsense knowledge graph for the αNLG task. Paul and Frank (2020) have proposed a multi-head knowledge attention method to incorporate commonsense knowledge to tackle the αNLI task. Unlike our previous work in Paul and Frank (2020), which focused on leveraging structured knowledge, in this work, we focus on learning about what will happen next from different counterfactual situations in a story context through language model fine-tuning. Specifically, we study the impact of such forward inference on the αNLI task in a multi-task learning framework and show how it can improve performance over a strong BERT model.

8 Conclusion

We have introduced a novel method for addressing the abductive reasoning task by explicitly learning what events could follow other events in a hypothetical scenario, and learning to generate such events, conditioned on a premise or hypothesis. We show how a language model – fine-tuned for this capability on a suitable narrative dataset – can be leveraged to support abductive reasoning in the αNLI tasks, in two settings: an unsupervised setting in combination with *BertScore*, to select the proper hypothesis, and a supervised setting in a $\mathcal{MTL}$ setting.

The relatively strong performance of our proposed models demonstrates that learning to choose from generated hypothetical next events the one that is most similar to the observation, supports the prediction of the most plausible hypothesis. Our experiments show that our unsupervised $LM_{\mathcal{I}}$+*BERTScore* model outperforms some of the strong supervised baseline systems on αNLI. Our research thus offers new perspectives for training generative models in different ways for various complex reasoning tasks.

Acknowledgements

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1. We thank our annotators for their valuable annotations. We also thank NVIDIA Corporation for donating GPUs used in this research.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

M. Bhatt and G. Flanagan. 2010. Spatio-temporal abduction for scenario and narrative completion (a preliminary statement). In *ECAI*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Igor Douven. 2017. Abduction. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2017 edition. Metaphysics Research Lab, Stanford University.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112.

A. Hayes and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1:77 – 89.

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, Hong Kong, China. Association for Computational Linguistics.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736, Online. Association for Computational Linguistics.

Antonis C Kakas, Robert A. Kowalski, and Francesca Toni. 1992. Abductive logic programming. *Journal of logic and computation*, 2(6):719–770.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When Choosing Plausible Alternatives, Clever Hans can be Clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Theo AF Kuipers. 1992. Naive and refined truth approximation. *Synthese*, 93(3):299–341.

Theo AF Kuipers. 2013. *From instrumentalism to constructive realism: On some relations between confirmation, empirical progress, and truth approximation*, volume 287. Springer Science & Business Media.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *arXiv preprint arXiv:2004.05569*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for 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 2822–2832.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

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

Dong Huk Park, L. Hendricks, Zeynep Akata, Anna Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8779–8788.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3671–3681, Minneapolis, Minnesota. Association for Computational Linguistics.

Debjit Paul and Anette Frank. 2020. Social commonsense reasoning with multi-head knowledge attention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2969–2980, Online. Association for Computational Linguistics.

Debjit Paul and Anette Frank. 2021. COINS: Dynamically Generating COntextualized Inference Rules for Narrative Story Completion. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, Online. Association for Computational Linguistics. Long Paper.

Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. 2020. Argumentative Relation Classification with Background Knowledge. In *Proceedings of the 8th International Conference on Computational Models of Argument (COMMA 2020)*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 319–330. Computational Models of Argument.

Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., CA.

C. S. Peirce. 1903. *Pragmatism as the Logic of Abduction*. https://www.textlog.de/7663.html.

Charles Sanders Peirce. 1965a. *Collected papers of Charles Sanders Peirce*, volume 5. Harvard University Press. http://www.hup.harvard.edu/catalog.php?isbn=9780674138001.

Charles Sanders Peirce. 1965b. *Pragmatism and pragmaticism*, volume 5. Belknap Press of Harvard University Press. https://www.jstor.org/stable/224970.

Harry E Pople. 1973. On the mechanization of abductive logic. In *Proceedings of the 3rd international joint conference on Artificial intelligence*, pages 147–152.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *2019 Conference on Empirical Methods in Natural Language Processing.*, Hongkong, China. Association for Computational Linguistics.

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

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. Ph.D. thesis, NUI Galway.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Parag Singla and Raymond J Mooney. 2011. Abductive markov logic for plan recognition. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Christian Strasser and G. Aldo Antonelli. 2019. Nonmonotonic logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2019 edition. Metaphysics Research Lab, Stanford University.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. 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 6076–6085, Hong Kong, China. Association for Computational Linguistics.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. ArXiv:2102.12060.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Yunchang Zhu, Liang Pang, Yanyan Lan, and Xueqi Cheng. 2020. $l2r^2$: Leveraging ranking for abductive reasoning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1961–1964, New York, NY, USA. Association for Computing Machinery.