Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision

Faeze Brahman[1]*, Vered Shwartz[2,3], Rachel Rudinger[4], and Yejin Choi[2,3]

[1]University of California Santa Cruz
[2]Allen Institute for AI
[3]Paul G. Allen School of Computer Science & Engineering, University of Washington
[4]University of Maryland, College Park, MD
fbrahman@ucsc.edu, rudinger@umd.edu, {vereds,yejinc}@allenai.org

Abstract

The black-box nature of neural models has motivated a line of research that aims to generate natural language rationales to explain why a model made certain predictions. Such rationale generation models, to date, have been trained on dataset-specific crowdsourced rationales, but this approach is costly and is not generalizable to new tasks and domains. In this paper, we investigate the extent to which neural models can reason about natural language rationales that explain model predictions, relying only on distant supervision with no additional annotation cost for human-written rationales. We investigate multiple ways to automatically generate rationales using pre-trained language models, neural knowledge models, and distant supervision from related tasks, and train generative models capable of composing explanatory rationales for unseen instances. We demonstrate our approach on the defeasible inference task, a nonmonotonic reasoning task in which an inference may be strengthened or weakened when new information (an update) is introduced. Our model shows promises at generating post-hoc rationales explaining why an inference is more or less likely given the additional information, however, it mostly generates trivial rationales reflecting the fundamental limitations of neural language models. Conversely, the more realistic setup of jointly predicting the update or its type and generating rationale is more challenging, suggesting an important future direction.

Introduction

Deep neural models perform increasingly well across NLP tasks, but due to their black-box nature, their success comes at the cost of our understanding of the system. The lack of transparency for *why* a model made a particular prediction may—among other problems—introduce fairness issues (Dodge et al. 2019), and hide the fact that often a model is right for the wrong reasons due to learning dataset-specific shortcuts and annotation artifacts (Gururangan et al. 2018; Poliak et al. 2018). There is growing interest in NLP in opening the black-box, through surrogate models (Ribeiro, Singh, and Guestrin 2016), counterfactual evaluation (Tenney et al. 2020), examining the inner structure of the neural network (Raffel et al. 2017; Jain et al. 2020), or generating natural language explanations. We focus on the latter approach. Recent work by Camburu et al. (2018) and Rajani et al. (2019) collected human-written explanations for the natural language inference (NLI; Bowman et al. 2015) and commonsense question answering (CommonSenseQA; Talmor et al. 2019) tasks and trained models to predict explanations for new instances. Such supervision is not always accessible, is expensive to obtain, and is unlikely to generalize well across datasets.

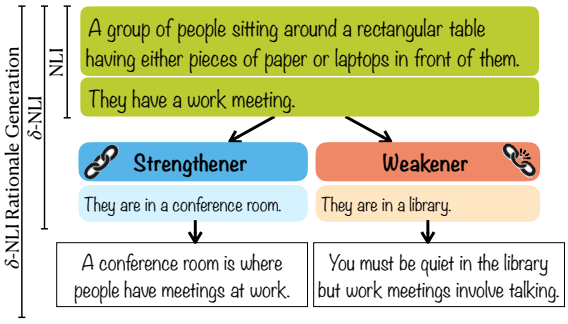

Figure 1: An illustration of the NLI, δ-NLI (Defeasible NLI), and δ-NLI Rationale Generation tasks.

In this work, we explore learning to rationalize using a distant supervision approach *without additional annotation cost*. We focus on the Defeasible Inference task (δ-NLI; Rudinger et al. 2020), illustrated in Figure 1. Given premise and hypothesis sentences, and an update sentence, the goal of the (discriminative) δ-NLI task is to recognize whether the update weakens or strengthens the entailment of the hypothesis by the premise. For example, the update that the people are in a conference room strengthens the hypothesis that they are in a work meeting. An alternative (generative) task is to generate the update given the premise, hypothesis, and update type (strengthener or weakener).

We present the Defeasible Inference Rationale Generation task, with the goal of generating natural language rationales that explain why a hypothesis is *more* likely after learning about a *strengthener* update and *less* likely after learning about a *weakener* update. To that end, we create the e-δ-NLI dataset by augmenting the δ-NLI dataset with rationales from various sources, including pre-trained language models, knowledge bases, and supervision from a related task. We then train two types of language model-based ra-

*Work done during internship at AI2.

tionale generation models: post-hoc models that generate a rationale given access the target values (i.e., the update or update type); and joint models that jointly generate the target value along with the rationale. The overall workflow of our approach is shown in Figure 2.

We evaluate the models with both automatic and human evaluations. The results of the post-hoc models are promising, with most generated rationales considered relevant and factually correct, and 40% on average considered explanatory. In line with prior work by Kumar and Talukdar (2020), further analysis revealed that models trained to post-hoc rationalize develop strategies to trivially map the target value to one of several patterns associated with it in the training data, such as "the update implies that hypothesis".

We consider the joint setup, in which the model has no access to the target value, to be more realistic. On this challenging setup, that hinders the models' ability to learn trivial shortcuts, the performance is worse, warranting future research in this direction.[1]

Background

Natural Language Inference. Recognizing Textual Entailment (RTE; Dagan et al. 2013), or, in its newer variant, Natural Language Inference (NLI; Bowman et al. 2015), is defined as a 3-way classification task. Given a premise sentence $\mathcal{P}$ and a hypothesis sentence $\mathcal{H}$, the goal is to determine whether $\mathcal{P}$ entails, contradicts, or is neutral with $\mathcal{H}$. $\mathcal{P}$ is said to entail $\mathcal{H}$ if a human reading $\mathcal{P}$ would typically infer that $\mathcal{H}$ is most likely true. $\mathcal{H}$ is neutral if it could be but is not necessarily true given $\mathcal{P}$.

In recent years, several large-scale datasets for the task have been released (e.g. Williams, Nangia, and Bowman 2018; Nie et al. 2020), encouraging training neural models. We focus on the Stanford Natural Language Inference dataset (SNLI; Bowman et al. 2015), in which image captions serve as premises, and hypotheses were crowdsourced.

Explainable NLI. Since deep learning has become the dominant paradigm in NLP research, efforts have been devoted to opening the "black-box" and interpreting neural models' predictions. One approach looks into the model's weights and traces back salient spans from the input that affected the prediction. The attention mechanism (Bahdanau, Cho, and Bengio 2015), which is popular across NLP models, facilitates this through the attention weights (Raffel et al. 2017; Jain et al. 2020). However, whether or not attention weights provide reliable insights into the model's decision-making process is debatable (Serrano and Smith 2019; Jain and Wallace 2019; Wiegreffe and Pinter 2019).

An alternative approach is to generate natural language explanations for the model's decision. This is typically done by training a model on free-form human explanations (Camburu et al. 2018; Rajani et al. 2019; Wang et al. 2019; Zellers et al. 2019), however, such supervision is not always available, and is costly to obtain. To that end, we propose a distant supervision approach that requires no additional supervision. Among other data source, we leverage the e-SNLI dataset (Camburu et al. 2018), in which premise-hypothesis pairs from SNLI have been augmented with human-written explanations for the gold labels.

There are several setups for interpretation methods: (i) **ante-hoc**: generating the rationale from the input, and providing it to the decision-making model with the input (Lei, Barzilay, and Jaakkola 2016; Bastings, Aziz, and Titov 2019; Kumar and Talukdar 2020) or without it (Jain et al. 2020); (ii) **joint**: generating the rationale and the label jointly (Narang et al. 2020); and (iii) **post-hoc**: generating a rationale given the input and the gold or predicted label. The motivation for the first approach is to produce faithful rationales, i.e. rationales representing the model's true decision process. However, there is no guarantee that the decision-making model actually uses the rationales. Moreover, in some cases the selected rationale is not sufficient to make the prediction without the input (Wiegreffe, Sarah and Marasović, Ana, and Smith, Noah A. 2020), while in others, label-specific rationale templates may make the label prediction trivial given the rationale (Kumar and Talukdar 2020). We focus on the latter two approaches: joint and post-hoc, while acknowledging that our rationales are not constructed to be faithful.[2]

Defeasible Inference. Defeasible reasoning is a non-monotonic logic in which valid inferences can become invalid when new information is introduced. For example, "Tweety is a bird" entails that "Tweety flies" unless provided with additional information such as "Tweety is a penguin" (Reiter 1980). Despite being a fundamental mode of human reasoning, modern NLP research paid little attention to non-monotonic reasoning (e.g. Qin et al. 2019; Bhagavatula et al. 2019). Recently, Rudinger et al. (2020) coupled defeasible reasoning with natural language inference by adding an update sentence $\mathcal{U}$ to the premise $\mathcal{P}$ and hypothesis $\mathcal{H}$. Expanding the traditional definition, $\mathcal{U}$ may either *weaken* or *strengthen* $\mathcal{H}$.

Two defeasible inference (δ-NLI) tasks were introduced: *discriminative defeasible inference*, in which given $\mathcal{P}$, $\mathcal{H}$, and $\mathcal{U}$, the goal is to classify the update as either weakener or strengthener (update type, $\mathcal{T}$); and *generative defeasible inference*, in which given $\mathcal{P}$, $\mathcal{H}$, and $\mathcal{T}$ the goal is to generate an update with the required type. The dataset for these tasks was built by crowdsourcing update sentences for neutral sentence-pairs from existing NLI datasets, Specifically, we use the SNLI portion of their data.

Unsupervised Knowledge Extraction from Pre-trained LMs. Pre-trained Language Models (LMs) based on the neural transformer architecture (Vaswani et al. 2017), such as GPT2 (Radford et al. 2019) and BERT (Devlin et al.

[1]The code and data are available at: https://github.com/fabrahman/RationaleGen.

[2]Humans also post-hoc rationalize decisions, and it is known to be flawed (Gazzaniga and LeDoux 2013). For recent works discussing rationale faithfulness, see Hase and Bansal (2020), Jacovi and Goldberg (2020), and Wiegreffe, Sarah and Marasović, Ana, and Smith, Noah A. (2020).

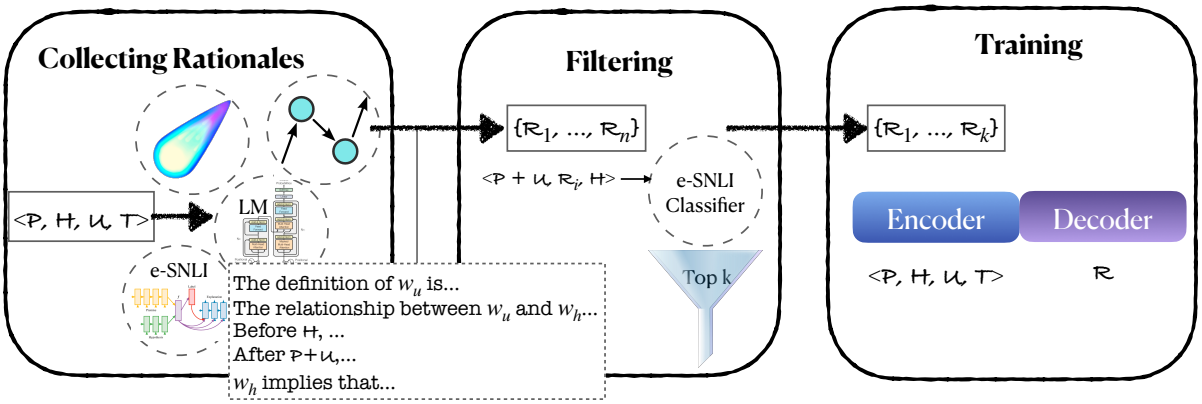

Figure 2: The complete training process: (1) collecting rationales from various sources, (2) Keeping the top k most helpful rationales; (3) training a generative model. During inference, we apply the generative model directly to the inputs.

2019) have greatly improved the performance on NLP tasks that require world knowledge and commonsense reasoning. While the best practice is to fine-tune the LM, they may also be used in an unsupervised manner. Petroni et al. (2019) and Davison, Feldman, and Rush (2019) completed commonsense knowledge bases (KB) by converting triplets into free-form text and predicting or scoring the target concept. Tamborrino et al. (2020) leveraged masked LMs to score the plausibility of answer choices in multiple-choice commonsense question answering (QA) tasks. Shwartz et al. (2020) used LMs to generate information-seeking clarification questions (e.g. "What is the definition of...") and their answers for providing relevant knowledge for commonsense QA tasks, which yielded similar performance gains to models utilizing KBs. Similarly, Latcinnik and Berant (2020) used LMs to generate a textual hypothesis which was used by the answer scorer of a multiple choice QA task.

e-δ-NLI Dataset

We describe e-δ-NLI (Explanations for Defeasible NLI). We augmented the δ-NLI dataset described in § with rationales that explain why a hypothesis is *more* likely after learning about a *strengthener* update and *less* likely after learning about a *weakener*. Rather than eliciting rationales from humans, we take a distant supervision approach and gather rationales from various sources, as exemplified in Table 1.

Collecting Rationales

Certain spans in the inputs $\mathcal{H}$ and $\mathcal{U}$ are highly salient for classifying the update type in the discriminative δ-NLI task. We hypothesize that these same spans will be salient for the task of generating rationales. Therefore we use the δ-NLI update type classifier and score each token in the input by its attention weight from the `<cls>` token in the final layer, and extract the set of top 20% non-continuous spans with respect to that score, denoted as S. For example, in Table 1, the most salient spans are underlined (hypothesis) and made italic (update). We use the following sources to extract or generate rationales.

Vanilla LM. We generate two types of rationales: definitions and purposes for single spans, and relationships for a pair of spans. We use SpaCy (Honnibal and Montani 2017) to keep only the grammatical salient spans $S_G \subseteq S$ by filtering out stop words and keeping both the entire (noun or verb) phrase and its head for each span.

Following Shwartz et al. (2020), we prompt the LM with "[context]. The definition of *np* is" for each noun phrase in S_G, and "[context]. The purpose of *vp* is" for verb phrases in S_G. We set the context to the concatenation of premise and hypothesis ($\mathcal{P}$+ $\mathcal{H}$) when the target phrase is in the hypothesis, and to $\mathcal{P}$+ $\mathcal{U}$ when it is from the update.

In addition, we generate the relationship between pairs of spans. We take the top 3 most similar pairs of s_u (subset of S_G originated from $\mathcal{U}$) and s_h (subset of S_G originated from $\mathcal{H}$), judged by the cosine similarity between their word2vec embeddings (Mikolov et al. 2013).[3] We prompt the LM with "$\mathcal{P}$+ $\mathcal{U}$+ $\mathcal{H}$. The relationship between s_u and s_h is that".

We use GPT2-M (Radford et al. 2019) via the Transformers package (Wolf et al. 2019). We limit the rationale length to up to 12 tokens, and use Nucleus sampling (Holtzman et al. 2020) with $p = 0.35$, and temperature $= 1.0$ to generate at most 20 rationales for each prompt.[4]

Knowledge-Enhanced LM. To further instill commonsense knowledge into the LM, we follow Guan et al. (2020) and continue pre-training GPT2-M on triplets from ConceptNet (Speer, Chin, and Havasi 2017) converted to natural language using the templates from Davison, Feldman, and Rush (2019). For example, `(a glass of milk, UsedFor, drinking)` is converted to "A glass of milk is used for drinking". We train the LM on the transformed triplets for 2 epochs. We then use the LM as previously detailed to generate definitions, purposes, and relationships. We use Nucleus sampling with $p = 0.5$, temperature $= 0.7$, and generate up to 5 rationales for each prompt.

[3] For multi-word spans we use maximum word-level similarity.

[4] Hyper-parameter values were chosen empirically from $p \in \{0.35, 0.5, 0.75\}$, temperature $\in \{0.7, 1\}$, #samples $\in \{5, 20\}$.

Source	Instance	Rationales
Vanilla LM	$\mathcal{P}$: [...] pedestrians walking down street filled with vendors and umbrella carts. $\mathcal{H}$: The vendors are there for the weekly farmer's market. **W**: They are *on* a busy *Manhattan* sidewalk selling *hotdogs.*	The relationship between "a busy Manhattan sidewalk selling hotdogs" and "weekly farmer's market" is that they both exist in tandem, but not necessarily together.
KG-Enhanced LM	$\mathcal{P}$: A person wearing red and white climbs a foggy mountain. $\mathcal{H}$: A person is rock climbing. **S**: The person is attached to a *rope going up* the side of the *mountain.*	The purpose of "rock climbing" is to reach a high place. The relationship between "rope" and "climbing" is that rope has property used to climb.
COMeT	$\mathcal{P}$: A baby boy in an elmo chair with lots of toys in the background. $\mathcal{H}$: The baby boy in the elmo chair is happy. **W**: The baby boy's mom is wiping tears from his eyes.	$\mathcal{H}$ precondition: The baby boy is seen as joyful. $\mathcal{U}$ postconditions: As a result, boy's mom feels to console.
NLI-derived	$\mathcal{P}$: The brown dog catches a ball in the air. $\mathcal{H}$: The dog plays with the ball outside. **S**: The ball skips into the bushes.	Catching a ball in the air implies that the dog plays with the ball. Bushes are outside.
NLI-derived w/ Highlights	$\mathcal{P}$: A woman wearing [...] and sunglasses, walks through a shopping outlet. $\mathcal{H}$: The woman is buying goods. **S**: The woman *is carrying* shopping *bags.*	If a woman is carrying bags, then she is buying goods.

Table 1: Examples of rationales generated from each of the sources. W stands for a weakener update and S for strengthener. Underline and *italic* show salient spans for hypothesis and update.

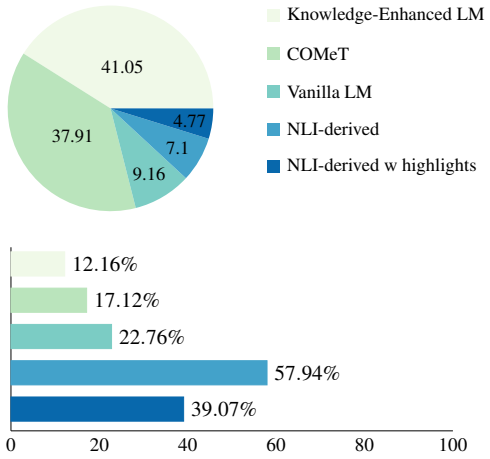

Figure 3: Top: Percentage of each source among the rationales in the final e-δ-NLI dataset. Bottom: Percentage of rationales that remained after filtering from each source.

COMeT. COMeT (Bosselut et al. 2019) is a LM-based knowledge base completion model. We use the model trained on ATOMIC (Sap et al. 2019), a commonsense KB consisting of *if-then* triplets concerning everyday situations, along multiple dimensions. We generate the postconditions following the update (`xWant`, `xEffect`, `xReact`, `xAttr`, `oWant`, `oEffect`, `oReact`) and the preconditions that lead to the hypothesis (`xNeed`, `xIntent`, `xAttr`). We use beam search with beam size of 5 as the decoding strategy, keeping the entire beam, and replace `PersonX` with the syntactic subject of the input sentence.

NLI-derived. We repurpose a model for the related task of NLI rationale generation for our task of rationale generation for δ-NLI. To that end, we reproduced the WT5 model suggested by Narang et al. (2020). The model is based on the T5 encoder-decoder language model (Raffel et al. 2020), and is trained on the e-SNLI dataset (Camburu et al. 2018) to jointly generate the label (entailment, contradiction) and the rationale for a given premise and hypothesis pair. More concretely, the input consists of the task prefix and the inputs (`explain nli premise:` $\mathcal{P}$ `hypothesis:` $\mathcal{H}$) while the expected output is `label explanation:` $\mathcal{R}$. During inference, we set the premise to $\mathcal{P}$+ $\mathcal{U}$, i.e., treating the update as part of the premise, and provide it to the model along with $\mathcal{H}$. The model generates the binary entailment label (excluding neutral) between $\mathcal{P}$+ $\mathcal{U}$ and $\mathcal{H}$, and the rationale that explains the label.[5]

NLI-derived with Highlights. Each instance in e-SNLI highlights salient spans in the input that the annotators considered helpful for explaining the label. We train a variant of the T5-based e-SNLI model that gets (only) the highlighted words as input and outputs the label and the rationale. We then generate rationales for the δ-NLI dataset by applying the model to salient spans in S that originated in $\mathcal{U}$ or $\mathcal{H}$.

[5]In practice, we only take the rationales and ignore the labels. But if we map entailment to strengthener and contradiction to weakener, we get 64% accuracy on the update type prediction. We note that this is an approximation. The definition of defeasible inference requires that a weakener makes the hypothesis *less* likely but not necessarily *unlikely*, while a strengthener makes the hypothesis *more* likely but not necessarily *likely*.

Task	Objective
Post-hoc Rationalization	
(1) Rationale	$P(\mathcal{R}\mid \mathcal{P}, \mathcal{H}, \mathcal{T}, \mathcal{U})$
(2) Update type	$P(\mathcal{T}\mid \mathcal{P}, \mathcal{H}, \mathcal{U}, \mathcal{R})$
(3) Update	$P(\mathcal{U}\mid \mathcal{P}, \mathcal{H}, \mathcal{T}, \mathcal{R})$
(4) Multi	(1) + (2) + (3)
Joint Prediction and Rationalization	
(5) Update type + Rationale	$P(\mathcal{T}, \mathcal{R}\mid \mathcal{P}, \mathcal{H}, \mathcal{U})$
(6) Update + Rationale	$P(\mathcal{U}, \mathcal{R}\mid \mathcal{P}, \mathcal{H}, \mathcal{T})$

Table 2: Different training setups we experiment with. We add special tokens to mark the boundaries of each input and output span (See Table 7 in Appendix).

Filtering Rationales

Following the collection step, each instance in the δ-NLI dataset is now augmented with a list of candidate rationales explaining its label (update type). To further improve the quality of this distant supervision, we rank and keep the best rationales. In particular, we would like to keep the rationales that are most helpful for predicting the label. Ideally, we would want to train a δ-NLI classifier that gets $\mathcal{P}$, $\mathcal{H}$, $\mathcal{U}$, and the rationale as input and outputs the update type. However, this causes a circular problem because we don't yet know which rationales are reliable.

Hence, again we use e-SNLI as a proxy. We train a classifier on e-SNLI that gets the premise, hypothesis, and rationale as inputs and predicts the entailment label (entailment, contradiction). Specifically, we fine-tune a binary RoBERTa classifier (Liu et al. 2019) with the following input format: $\mathcal{P}$ <sep> $\mathcal{R}$ <sep> $\mathcal{H}$. For a δ-NLI instance ($\mathcal{P}$, $\mathcal{H}$, $\mathcal{T}$, $\mathcal{U}$) with a set of candidate rationales $\{\mathcal{R}_i\}_{i=1}^{N_R}$ (of various sources), we compute: $o = \text{NLI}(\mathcal{P} + \mathcal{U}$ <sep> $\mathcal{R}_i$<sep> $\mathcal{H})$, where o is a 2-dimensional vector representing the confidence of the classifier in each label. We score each rationale by the confidence assigned to the label associated with its update type: strengtheners as entailment and weakener[6] as contradiction, and rank the rationales accordingly. We keep the top 10% ranked rationales for each instance, yielding 8 rationales per instance on average.

We follow the original split to train (80%), test (10%), and development (10%) sets. By augmenting the data with multiple rationales per original δ-NLI instance, the final e-δ-NLI dataset consists of 731,579 training, 15,781 test, and 15,527 development instances. Figure 3 shows the percent of rationale sources in the dataset.

Rationale Generation Model

We use the e-δ-NLI dataset to train various generative models with the goal of generating rationales that explain why a hypothesis is *more* likely after learning about a *strengthener* update and *less* likely after learning about a *weakener*.

Every instance in the e-δ-NLI dataset consists of a premise $\mathcal{P}$, hypothesis $\mathcal{H}$, update type $\mathcal{T}$, update $\mathcal{U}$, and a set of rationales $\{\mathcal{R}_i\}_{i=1}^{N_R}$. During training, we treat every ($\mathcal{P}$, $\mathcal{H}$, $\mathcal{T}$, $\mathcal{U}$, $\mathcal{R}$) for $\mathcal{R} \in \{\mathcal{R}_i\}_{i=1}^{N_R}$ as a separate instance.

Architecture and Implementation Details

We fine-tune transformer-based pre-trained LMs on the e-δ-NLI dataset. Specifically, we use GPT2-XL (Radford et al. 2019) and Bart-L (Lewis et al. 2020).[7] We use the Transformers package (Wolf et al. 2019), training each model for a single epoch with batch size of 8 (GPT2), and 128 (Bart) on a Quadro RTX 8000 GPU machine.

Training Objective

We minimize the conditional log-likelihood of the output given the input: $\mathcal{L} = -\sum_{i=1}^{n} \log p(x_i^{out} | x_{<i}^{out}, x^{in})$. In particular, for GPT2, which is a standard LM model, the loss is computed over the entire sequence $[x^{in}; x^{out}]$, whereas in Bart, which is an encoder-decoder model, the loss is computed only over the output sequence, x_{out}.

We experiment with various training setups described in Table 2. Our setups can be divided into two categories. The first category is **Post-hoc Rationalization**, in which the model has access to the target values (i.e., update and update type) and is required to explain it. Our main task in this category is Rationale Generation (1). It is formulated as generating a rationale conditioned on the premise, hypothesis, update, and update type. Similarly, we can generate each of the update type (2) and update (3) given all other fields. These two setups are orthogonal to our goal, but we combine them with (1) in a multi-task setup (4) where we expect them to improve the model's generalizability (Shwartz and Dagan 2018; Zellers et al. 2019) and improve the performance on the main task. The second and more realistic category is **Joint Prediction and Rationalization**, in which the model jointly predicts either update type (5) or update (6) along with an explanation.

Results

For each combination of rationale generation training setup, we generated a rationale for each instance in the test set using beam search with 5 beams. We evaluated the generated rationales both in terms of automatic metrics and human evaluation. The results are shown in Table 3.

Automatic Evaluation

We used standard n-gram overlap metrics: the precision-oriented BLEU score (Papineni et al. 2002) and recall-oriented ROUGE score (Lin 2004). Specifically, we used BLEU-4 that measures overlap of n-grams up to $n = 4$, and ROUGE-L that measures longest matching sequences, and compared multiple predictions against multiple distantly supervised rationales as references. The result of the automatic measures are reported in Table 3. In general, GPT2-based models achieve better automatic scores. We also observe additive gain using multi-task setup on both BLEU and ROUGE scores.

[6] Equating strengtheners with entailment and weakeners as contradiction is a simplifying assumption, which is not strictly true.

[7] In our preliminary experiments, we also experimented with T5, but we did not observe any improvements.

Objective	Model	Automatic		Human (%)			
		BLEU	ROUGE	Gram.	Rele.	Corr.	Expl.
Post-hoc Rationalization							
Rationale	GPT2-XL	33.0	33.91	92.5/93.5	33.5/60.0	52.5/45.5	2.0/4.5
	BART-L	13.15	22.48	**95.0/99.5**	**80.0/80.5**	**55.0/58.0**	**47.0/33.0**
Multi	GPT2-XL	**33.58**	**34.50**	88.5/94.5	30.0/55.0	47.0/43.0	0.5/7.0
	BART-L	17.38	24.03	95/97.5	74.5/76.5	55.5/53.0	44.0/22.5
Joint Prediction and Rationalization							
Update	GPT2-XL	23.93	**31.71**	85.0/88.0	15.5/35.0	30.5/15.5	1.0/1.5
+ Rationale	BART-L	25.24	30.83	86.5/83.0	20.0/34.5	34.0/17.5	2.5/0.5
Update type	GPT2-XL	**27.90**	31.18	86.5/89.0	**27.0/39.0**	**36.5/29.0**	**8.5/3.0**
+ Rationale	BART-L	24.54	29.04	86.5/85.5	26.0/18.5	35.5/30.5	7.0/1.0

Table 3: Automatic and human evaluation of rationale generation for the test set. Human evaluation results are presented for strengtheners and weakeners separately (S/W).

Human Evaluation

Since automatic metrics have demonstrated low correlation with human judgments across various NLG tasks (Novikova et al. 2017), and because our automatic metrics only evaluate the generated rationales against the distantly supervised rationales (in place of human-written references), we also conduct a more reliable evaluation using human judges on Amazon Mechanical Turk. We sampled 200 instances, along with a generated rationale for each model. Following Shwartz et al. (2020), we asked workers to determine whether a rationale was 1) grammatical, not entirely grammatical but understandable, or completely not understandable; 2) relevant to the instance ($\mathcal{P}$, $\mathcal{H}$, and $\mathcal{U}$); 3) factually correct or likely true; and 4) explanatory of the update type (i.e. why the strengthener makes the hypothesis more likely or the weakener makes it less likely). To ensure the quality of annotations, we required that the workers be located in the US, UK, or Canada, and have a 99% approval rate for at least 5,000 prior tasks. We aggregated annotations from 3 workers using majority vote. The annotations yielded fair levels of agreement, with Fleiss' Kappa (Landis and Koch 1977) between $\kappa = 0.22$ for relevance and $\kappa = 0.37$ for being explanatory. We analyze the results from the following perspectives:

Best Setup. Across models, most rationales are grammatical or understandable (83%-99%). The best performance is achieved by Rationale BART-L, in which 80% of the rationales were considered relevant, over 55% correct, and between 33% (weakeners) to 47% (strengtheners) explanatory. Also, in general, better rationales are generated for strengthener than weakener.

LM and Objective. The multi-task setup did not improve the rationale generation performance. Among the post-hoc rationalization category, Bart-based models substantially outperformed GPT2-based models.

Post-hoc vs. Joint. In the post-hoc rationalization setups, access to the target values (update more than update type) yielded more explanatory rationales (Expl. score in Table 3), but as discussed later, they are often trivial. The joint setup proved to be extremely challenging, with only 0.5%-8.5% of the rationales considered explanatory.

Pattern	%
Strengtheners	
[S] ([H]) implies (that) [H] ([S])	64.9
[S] ([H]) is a rephrasing of [H] ([S])	14.9
[H] ([S]) because [S] ([H])	12.8
[S] means [H]	2.1
[S] is [H]	1.1
[S] is the same as [H]	1.1
Other	3.19
Weakeners	
Something cannot be [W] and [H] at the same time	33.3
Something cannot be [W] ([H]) if it is [H] ([W])	31.8
[W] is not the same as [H]	13.6
Something is either [W] or [H]	10.6
[W] is not [H]	6.1
Other	4.6

Table 4: Patterns of rationales generated by Rationale Bart-L that were considered explanatory. H, S, and W stand for Hypothesis, Strengthener and Weakener.

Analysis

Quality of the Distant Supervision

We study the quality of rationales in the e-δ-NLI dataset through human evaluation. We repeated the same crowdsourcing setup, this time evaluating the distantly supervised rationales (i.e. after filtering) of 100 random instances.

Table 8 in Appendix shows that the quality of the training data is surprisingly worse than that of the generated rationales. Specifically, rationales originating from LMs are often judged as incorrect and non-explanatory, much due to statements such as "The definition of s is s". Conversely, NLI-derived rationales are identified as the most explanatory ones, in agreement with our filtering step which kept

Error Type	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Strengthener	4	20	0	16	22	18	6
Weakener	44	14	28	8	0	0	4
Overall	24	17	14	12	11	9	5

Table 5: Percent of rationales with each error type.

the highest percents of NLI-derived rationales (58%). As we show in previous section, most generated rationales are in the format of e-SNLI rationales, which might explain the discrepancy between the quality of the generated rationales and that of the training data (in which only 7.1% of the rationales are NLI-derived rationales).

Quality of Generated Rationales

We manually analyzed the rationales generated by the best model (Rationale Bart-L) that were considered grammatical, relevant, and correct by humans.

Explanatory. We analyzed the 160 rationales that were considered "explanatory" (94 for strengtheners and 66 for weakeners), and found that almost all of them fit into one of several patterns of rationales that are trivial to generate given the target value (update type). These patterns are displayed in Table 4. We see this as further motivation to focus on the joint setup in future research.

Non-Explanatory. We sampled and analyzed 100 rationales that were annotated as "non-explanatory" by workers (50-50 for strengtheners and weakeners). We found the following common types of errors, and categorized each rationale into one or more categories. The result is shown in Table 5, and exemplified in Table 9.

(1) **Insufficient**: providing one of several required reasoning hops.

(2) **Incorrect implications**: following one of the templates in Table 4, but not making sense.

(3) **Incorrect post/pre-conditions**: involving wrong inferences about the post-conditions of $\mathcal{U}$ or the pre-conditions of $\mathcal{H}$.

(4) **Partially correct**: following a pattern in Table 4, incorrectly using part of $\mathcal{U}$ or $\mathcal{H}$. For example in Table 9, "the group is on vacation is a rephrasing of resort", instead of rephrasing of "they are at a resort".

(5) **Repetitive statements**: defining terms or relationships between a pair of terms, by repeating the term ("The definition of s is s").

(6) **Wrong template**: following wrong templates in Table 4, e.g. generating "X is a rephrasing of Y" when X implies Y ("The people are eating fresh seafood is a rephrasing of sitting near the ocean").

(7) **Rationalizing the premise**: the rationale explains the premise instead of the hypothesis (e.g. "$\mathcal{U}$ implies $\mathcal{P}$").

Model	Gram.	Rele.	Corr.	Expl.
Rationale BART-L	95/99.5	80/80.5	55/58	47/33
w/o Filtering	99/100	94/93.5	39.5/25	49/26
NLI-derived only	99.5/97	100/99	52.5/32.5	50.5/29

Table 6: Ablation studies human evaluation. Results (percents) are presented for strengtheners and weakeners (S/W).

We observe a large portion of errors (especially for weakener) are from error type (1) where the rationale needs to be completed by another hop of reasoning.

Ablation Studies

We conduct ablation studies in which we ablate either (i) the filtering step (randomly selecting a rationale from each source), or (ii) all sources besides NLI-derived rationales from our e-δ-NLI dataset. In both cases, we trained the best setup (Rationale Bart-L) and evaluated the results using the same human evaluation setup described earlier.

The results are reported in Table 6. Both ablations increase the relevance of rationales while hurting their factual correctness and producing less explanatory weakener rationales. In the case of the second ablation, this is likely due to the fact that most model-generated rationales in the format of the NLI-derived rationales copy parts of the input into label-specific templates, yielding relevant but not necessarily correct or explanatory rationales.

Conclusion

We presented an approach for generating rationales for the defeasible inference task, i.e., explaining why a given update either strengthened or weakened the hypothesis. We experimented with various training setups categorized into post-hoc rationalization and joint prediction and rationalization. Rather than collecting human explanations, we chose to train our models in a distant supervision approach that requires no additional annotation cost and may generalize better across datasets. The results indicated that the post-hoc rationalization setup is easier than the joint setup, with many of the post-hoc generated rationales considered by humans as explanatory. Nonetheless, the model's success may be attributed to its access to the update type, which enabled learning a trivial mapping from the update type to rationale templates associated with it in the training data. The joint setup, on the other hand, proved to be more challenging. We hope that future work will focus on jointly predicting a label and generating a rationale, which is a more realistic setup and which may yield less trivial and more faithful rationales.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. We also thank Ana Marasović, Sarah Wiegreffe, xlab and Mosaic team members for helpful discussions.

Training Setups and More Analysis

Input and output formats for each task are shown Table 7.

Task	Input	Output
	Post-hoc Rationalization	
(1)	[premise] $\mathcal{P}$ [hypo] $\mathcal{H}$ [ut] <$\mathcal{T}$> [update] $\mathcal{U}$ [rationale]	$\mathcal{R}$
(2)	[premise] $\mathcal{P}$ [hypo] $\mathcal{H}$ [update] $\mathcal{U}$ [rationale] $\mathcal{R}$ [ut]	<$\mathcal{T}$>
(3)	[premise] $\mathcal{P}$ [hypo] $\mathcal{H}$ [rationale] $\mathcal{R}$ [ut] <$\mathcal{T}$> [update]	$\mathcal{U}$
(4)	(1) + (2) + (3)	
	Joint Prediction and Rationalization	
(5)	[premise] $\mathcal{P}$ [hypo] $\mathcal{H}$ [update] $\mathcal{U}$ [update_type_rationale]	[ut] <$\mathcal{T}$> [rationale] $\mathcal{R}$
(6)	[premise] $\mathcal{P}$ [hypo] $\mathcal{H}$ [ut] <$\mathcal{T}$> [update_rationale]	[update] $\mathcal{U}$ [rationale] $\mathcal{R}$

Table 7: Input and output format for each tasks.

Source	Gram.	Rele.	Corr.	Expl.
Vanilla LM	6.54/73.03	4.67/14.61	3.74/4.49	0.93/2.25
K-Enhanced LM	86.01/70.27	35.42/17.57	32.14/22.97	2.08/2.70
COMeT	91.46/87.16	73.17/47.54	47.56/12.84	14.63/8.20
NLI-derived	97.56/100.0	97.56/91.11	70.73/44.44	63.41/64.44
NLI-derived w/ H	50.00/84.62	38.24/53.85	17.65/30.77	17.65/30.77
Overall	71.33/83.83	39.50/42.50	31.00/16.00	8.67/11.83

Table 8: Human evaluation for the distant supervision rationales in the test set. Results (percents) are presented for strengtheners and weakeners separately (S/W).

Error	Instance & Rationale
(1)	$\mathcal{P}$: Four individuals are sitting on a small dock by the water as a boat sails by. $\mathcal{H}$: Four people sitting near the ocean. **W**: They're in Egypt. $\mathcal{R}$: Before, four people needed to go to the beach.
(2)	$\mathcal{P}$: Two men in orange uniforms stand before a train and do some work. $\mathcal{H}$: Tall humans working. **S**: The men can easily touch the top of the train with their hands. $\mathcal{R}$: The men can [...] train with their hands implies that they are working.
(3)	$\mathcal{P}$: A cyclist dressed in black and white is pointing. $\mathcal{H}$: A cyclist dressed in black and white points towards the sky. **W**: A man asked the cyclist which building is the bank. $\mathcal{R}$: Before, a cyclist needed to go to the store.
(4)	$\mathcal{P}$: A group of people are sight seeing and taking pictures. $\mathcal{H}$: the group is on vacation. **S**: They are at a resort. $\mathcal{R}$: The group is on vacation is a rephrasing of resort.

Table 9: Examples for the common error types.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

Bastings, J.; Aziz, W.; and Titov, I. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. In *ACL*.

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

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

Bowman, S.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Neurips*.

Dagan, I.; Roth, D.; Sammons, M.; and Zanzotto, F. M. 2013. *Recognizing Textual Entailment: Models and Applications*. Morgan & Claypool Publishers.

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

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

Dodge, J.; Liao, Q. V.; Zhang, Y.; Bellamy, R. K. E.; and Dugan, C. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *IUI*.

Gazzaniga, M. S.; and LeDoux, J. E. 2013. *The integrated mind*. Springer Science & Business Media.

Guan, J.; Huang, F.; Zhao, Z.; Zhu, X.; and Huang, M. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *TACL* .

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL*.

Hase, P.; and Bansal, M. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In *ACL*.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *ICLR*.

Honnibal, M.; and Montani, I. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear* 7(1).

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *ACL*.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *NAACL*.

Jain, S.; Wiegreffe, S.; Pinter, Y.; and Wallace, B. C. 2020. Learning to Faithfully Rationalize by Construction. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *ACL*.

Kumar, S.; and Talukdar, P. P. 2020. NILE : Natural Language Inference with Faithful Natural Language Explanations. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *ACL*.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics* .

Latcinnik, V.; and Berant, J. 2020. Explaining question answering models through text generation. *arXiv* .

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *EMNLP*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* .

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv* .

Narang, S.; Raffel, C.; Lee, K.; Roberts, A.; Fiedel, N.; and Malkan, K. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. *arXiv* .

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *ACL*.

Novikova, J.; Dušek, O.; Curry, A. C.; and Rieser, V. 2017. Why We Need New Evaluation Metrics for NLG. In *EMNLP*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *EMNLP-IJCNLP*. Hong Kong, China.

Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; and Van Durme, B. 2018. Hypothesis Only Baselines in Natural Language Inference. In **SEM*.

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

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

Raffel, C.; Luong, M.; Liu, P. J.; Weiss, R. J.; and Eck, D. 2017. Online and Linear-Time Attention by Enforcing Monotonic Alignments. In *CML*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR* .

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *ACL*.

Reiter, R. 1980. A logic for default reasoning. *Artificial intelligence* .

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *ACM SIGKDD*.

Rudinger, R.; Shwartz, V.; Hwang, J. D.; Bhagavatula, C.; Forbes, M.; Le Bras, R.; Smith, N. A.; and Choi, Y. 2020. Thinking Like a Skeptic: Defeasible Inference in Natural Language. In *Findings of ACL: EMNLP*.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In *ACL*.

Shwartz, V.; and Dagan, I. 2018. Paraphrase to Explicate: Revealing Implicit Noun-Compound Relations. In *ACL*.

Shwartz, V.; West, P.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2020. Unsupervised Commonsense Question Answering with Self-Talk. In *EMNLP*.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: an open multilingual graph of general knowledge. In *AAAI*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *NAACL*.

Tamborrino, A.; Pellicanò, N.; Pannier, B.; Voitot, P.; and Naudin, L. 2020. Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning. In *ACL*.

Tenney, I.; Wexler, J.; Bastings, J.; Bolukbasi, T.; Coenen, A.; Gehrmann, S.; Jiang, E.; Pushkarna, M.; Radebaugh, C.; Reif, E.; and Yuan, A. 2020. The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. In *EMNLP*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Neurips*.

Wang, C.; Liang, S.; Zhang, Y.; Li, X.; and Gao, T. 2019. Does it Make Sense? And Why? A Pilot Study for Sense Making and Explanation. In *ACL*.

Wiegreffe, S.; and Pinter, Y. 2019. Attention is not not Explanation. In *EMNLP-IJCNLP*.

Wiegreffe, Sarah and Marasović, Ana, and Smith, Noah A. 2020. Measuring Association Between Labels and Free-Text Rationales. *arXiv* .

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL-HLT*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* .

Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019. Defending Against Neural Fake News. In *Neurips*.