Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision

Faeze Brahman1*, Vered Shwartz2,3, Rachel Rudinger4, and Yejin Choi2,3

1University of California Santa Cruz
2Allen Institute for AI
3Paul G. Allen School of Computer Science & Engineering, University of Washington
4University 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 promise 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.

In this work, we explore learning to rationalize using a distant supervision approach without additional annotation cost. We focus on the Defeasible Inference task ($\delta$-NLI; Rudinger et al. 2020), illustrated in Figure 1. Given premise and hypothesis sentences, and an update sentence, the goal of the (discriminative) $\delta$-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-$\delta$-NLI dataset by augmenting the $\delta$-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 r-
tional generation models: post-hoc models that generate a rationale given access to 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 $P$ and a hypothesis sentence $H$, the goal is to determine whether $P$ entails, contradicts, or is neutral with $H$. $P$ is said to entail $H$ if a human reading $P$ would typically infer that $H$ is most likely true. $H$ is neutral if it could be but is not necessarily true given $P$.

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 diagnostic 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, Azz, 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 $U$ to the premise $P$ and hypothesis $H$. Expanding the traditional definition, $U$ may either weaken or strengthen $H$.

Two defeasible inference ($\delta$-NLI) tasks were introduced: discriminative defeasible inference, in which given $P$, $H$, and $U$, the goal is to classify the update as either weaker or stronger (update type, $T$); and generative defeasible inference, in which given $P$, $H$, and $T$ 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. 2019), are becoming increasingly popular. 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).

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

²Humans 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).
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. Schwartz et al. (2020) used LMs to generate information-seeking clarifications (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 *strengthening* update and less likely after learning about a *weakening* update. 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 $<\text{cls}>$ 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 Schwartz 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 ($P+H$) when the target phrase is in the hypothesis, and to $P+U$ 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 $s_a$ (subset of $S_G$ originated from $\mathcal{U}$) and $s_b$ (subset of $S_G$ originated from $\mathcal{H}$), judged by the cosine similarity between their word2vec embeddings (Mikolov et al. 2013). We prompt the LM with “$P+U+H$. The relationship between $s_a$ and $s_b$ 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.

---

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

4Hyper-parameter values were chosen empirically from $p \in \{0.35, 0.5, 0.75\}$, temperature $\in \{0.7, 1\}$, #samples $\in \{5, 20\}$.
The relationship between “rope” and “climbing” is that rope and climbing exist in tandem, but not necessarily together.

The purpose of “rock climbing” is to reach a high place.

We note that this is an approximation. The definition of defeasible labels. But if we map entailment to strengthener and contradiction to weaker, we get 64% accuracy on the update type prediction.

The vendors are there for the weekly farmer’s market.

Vanilla LM

<table>
<thead>
<tr>
<th>Source</th>
<th>Instance</th>
<th>Rationales</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The vendors are there for the weekly farmer’s market.</td>
<td>The vendors are there for the weekly farmer’s market.</td>
</tr>
<tr>
<td></td>
<td>They are on a busy Manhattan sidewalk selling hotdogs.</td>
<td>They are on a busy Manhattan sidewalk selling hotdogs.</td>
</tr>
</tbody>
</table>

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

**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 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 U or H. In practice, we only take the rationales and ignore the labels. But if we map entailment to strengthener and contradiction to weaker, we get 64% accuracy on the update type prediction. We note that this is an approximation. The definition of defeasible inference requires that a weaker makes the hypothesis less likely but not necessarily unlikely, while a strengthener makes the hypothesis more likely but not necessarily likely.
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 \( P, H, U, \) 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: \( P <\)sep\( > R <\)sep\( > H. \) For a δ-NLI instance (\( P, H, T, U \)) with a set of candidate rationales \( \{ R_i \}_{i=1}^{N_R} \) (of various sources), we compute: \( o = \text{NLI}(P + U <\)sep\( > R_i <\)sep\( > H), \) 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 it (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 \( P, H, U, \) 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.

Joint Prediction and Rationalization
We use the e-SNLI dataset to train various generative models with the goal of generating rationales that explain why a hypothesis is more likely after learning about a strengthenere update and less likely after learning about a weaker.

Every instance in the e-δ-NLI dataset consists of a premise \( P, \) hypothesis \( H, \) update type \( T, \) update \( U, \) and a set of rationales \( \{ R_i \}_{i=1}^{N_R}. \) During training, we treat every \( (P, H, T, U, R) \) for \( R \in \{ R_i \}_{i=1}^{N_R}, \) 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: \( \mathcal{L} = -\sum_{i=1}^{n} \log p(x^{out} | x^{in}, a). \) 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.
### 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).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT2-XL</td>
<td>33.0</td>
<td>33.91</td>
<td>92.5</td>
<td>53.5</td>
<td>52.5</td>
<td>2.0</td>
</tr>
<tr>
<td>BART-L</td>
<td>13.15</td>
<td>22.48</td>
<td>80.0</td>
<td>30.0</td>
<td>55.0</td>
<td>42.5</td>
</tr>
</tbody>
</table>

### 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 Schwartz 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 (P, H, and U); 3) factually correct or likely true; and 4) explanatory of the update type (i.e. why the strengtheners 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 strengtheners 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.

### Analysis

**Quality of the Distant Supervision**

We study the quality of rationales in the e-$\delta$-NLI dataset through human evaluation. We repeated the same crowd-sourcing 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
Rationalizing the premise

Wrong template

(6)

Repetitive statements

(5)

Non-Explanatory.

We sampled and analyzed 100 ratio-

Joint setup in future research.

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.

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 ratio-

nals that were annotated as “non-explanatory” by workers

(50-50 for strengtheners and weakeners). We found the fol-

lowing common types of errors, and categorized each ra-

tionale into one or more categories. The result is shown in

Table 5, and exemplified in Table 9.

(1) Insufficient: providing one of several required reason-

ing hops.

(2) Incorrect implications: following one of the templates

in Table 4, but not making sense.

(3) Incorrect post/pre-conditions: involving wrong in-

ferences about the post-conditions of \(U\) or the pre-

conditions of \(H\).

(4) Partially correct: following a pattern in Table 4, incor-

rectly using part of \(U\) or \(H\). 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 Ta-

ble 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. “\(U\) implies \(P\)”).

<table>
<thead>
<tr>
<th>Error Type</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strengtheners</td>
<td>4</td>
<td>20</td>
<td>0</td>
<td>16</td>
<td>22</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>Weakener</td>
<td>14</td>
<td>14</td>
<td>28</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Overall</td>
<td>24</td>
<td>17</td>
<td>14</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>5</td>
</tr>
</tbody>
</table>

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 ra-

tionales 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 ratio-

nals that were annotated as “non-explanatory” by workers

(50-50 for strengtheners and weakeners). We found the fol-

lowing common types of errors, and categorized each ra-

tionale into one or more categories. The result is shown in

Table 5, and exemplified in Table 9.

(1) Insufficient: providing one of several required reason-

ing hops.

(2) Incorrect implications: following one of the templates

in Table 4, but not making sense.

(3) Incorrect post/pre-conditions: involving wrong in-

ferences about the post-conditions of \(U\) or the pre-

conditions of \(H\).

(4) Partially correct: following a pattern in Table 4, incor-

rectly using part of \(U\) or \(H\). 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 Ta-

ble 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. “\(U\) implies \(P\)”).

We observe a large portion of errors (especially for weak-

ener) 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-\(\delta\)-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 in-
crease the relevance of rationales while hurting their factual

correctness and producing less explanatory weaker ration-

nals. 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 up-
date either strengthened or weakened the hypothesis. We
experimented with various training setups categorized into
post-hoc rationalization and joint prediction and rationaliza-
tion. 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 hu-
mans 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 la-
bel 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 com-
ments. 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.

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


Rationale BART-L 99/99.5 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 5: Percent of rationales with each error type.
Table 7: Input and output format for each tasks.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla LM</td>
<td>6.54/73.03</td>
<td>4.67/14.61</td>
<td>3.74/4.49</td>
<td>0.93/2.25</td>
</tr>
<tr>
<td>K-Enhanced LM</td>
<td>86.01/70.27</td>
<td>35.42/17.57</td>
<td>32.14/22.97</td>
<td>2.08/2.70</td>
</tr>
<tr>
<td>COMeT</td>
<td>91.46/87.16</td>
<td>73.17/47.54</td>
<td>47.56/12.84</td>
<td>14.63/6.20</td>
</tr>
<tr>
<td>NLI-derived</td>
<td>97.56/100.0</td>
<td>97.56/91.11</td>
<td>70.73/44.44</td>
<td>63.41/64.44</td>
</tr>
<tr>
<td>NLI-derived w/ H</td>
<td>50.00/84.62</td>
<td>38.24/53.85</td>
<td>17.65/16.00</td>
<td>17.65/64.44</td>
</tr>
<tr>
<td>Overall</td>
<td>71.33/83.83</td>
<td>39.50/42.50</td>
<td>31.00/16.00</td>
<td>8.67/11.83</td>
</tr>
</tbody>
</table>

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

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

Table 9: Examples for the common error types.

References


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


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


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.


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


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


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.


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


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


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


