Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi
AAAI 2021
Learning to **Rationalize** for **Nonmonotonic Reasoning** with **Distant Supervision**
Background

Opening the “black-box” and interpreting neural models’ predictions:

🤖 Surrogate models [Ribeiro et al. 2016]
Background

Opening the “black-box” and interpreting neural models’ predictions:

🤖 Surrogate models [Ribeiro et al. 2016]

🤖 Counterfactual evaluation [Tenney et al 2020]
Background

Opening the “black-box” and interpreting neural models’ predictions:

🤖 Surrogate models [Ribeiro et al. 2016]

🤖 Counterfactual evaluation [Tenney et al. 2020]

🤖 Examining inner structure of NN, attention weights [Collin et al 2017, Jain et al. 2020]
Background

Opening the “black-box” and interpreting neural models’ predictions:

🤖 Surrogate models [Ribeiro et al. 2016]

🤖 Counterfactual evaluation [Tenney et al 2020]

🤖 Examining inner structure of NN, attention weights [Collin et al 2017, Jain et al. 2020]

🤖 Generating natural language explanations for the model's decisions
Defeasible Inference

A nonmonotonic mode of reasoning in which an initial supported inference may be weakened or overturned in the light of new evidence!
Defeasible Inference

A nonmonotonic mode of reasoning in which an initial supported inference may be weakened or overturned in the light of new evidence!

P: Tweety is a bird.
A nonmonotonic mode of reasoning in which an initial supported inference may be weakened or overturned in the light of new evidence!

P: Tweety is a bird.
H: Tweety flies.
A nonmonotonic mode of reasoning in which an initial supported inference may be weakened or overturned in the light of new evidence!

P: Tweety is a bird.
H: Tweety flies.
U: Tweety is a penguin.
Given a premise $P$ and hypothesis $H$, an update $\nu$ is called:

$P$: Tweety is a bird.

$H$: Tweety flies.
Defeasible Inference

Given a premise $P$ and hypothesis $H$, an update $\nu$ is called:

**weakened** $\rightarrow$ if a human would most likely find $H$ *less likely to be true* after learning $\nu$;

$P$: Tweety is a bird.

$H$: Tweety flies.

Weakener: Tweety is a penguin.
Given a premise $P$ and hypothesis $H$, an update $\nu$ is called:

**weakened** $\rightarrow$ if a human would most likely find $H$ *less likely to be true* after learning $\nu$;

**strengthener** $\rightarrow$ if they would find $H$ *more likely to be true*

---

P: Tweety is a bird.
H: Tweety flies.

Weakener: Tweety is a penguin.
Strengthener: Tweety is on a tree.
Defeasible Inference

Discriminative Task

Given a premise $P$, a hypothesis $H$, and an update $U$, the goal is to predict the update type $T$, i.e. whether $U$ strengthens or weakens $H$. 

$P$ \quad \quad \quad T

$H$ \quad \quad U
Defeasible Inference

Given a premise \( \mathcal{P} \), a hypothesis \( \mathcal{H} \), and an update \( \mathcal{U} \), the goal is to predict the update type \( \mathcal{T} \), i.e. whether \( \mathcal{U} \) strengthens or weakens \( \mathcal{H} \).

\[
\begin{align*}
\mathcal{P} : & \quad \text{A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.} \\
\mathcal{H} : & \quad \text{They have a work meeting.} \\
\mathcal{U} : & \quad \text{They are in a conference room.} \\
\mathcal{T} : & \quad \text{Strengthener}
\end{align*}
\]
Defeasible Inference

Discriminative Task

Given a premise $P$, a hypothesis $H$, and an update $\nu$, the goal is to predict the update type $T$, i.e. whether $\nu$ strengthens or weakens $H$.

![Diagram](image)

Thinking Like a Skeptic: Defeasible Inference in Natural Language.
Given a premise \( \mathcal{P} \), a hypothesis \( \mathcal{H} \), and a desired update type \( \tau \) (\textit{weakener} or \textit{strengthener}), the goal is to generate an update \( \mathcal{U} \) that satisfies the type constraint.

**Generative Task**

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.  
They have a work meeting.  \textbf{Strengthener}

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.  
They are in a conference room.  
They have a work meeting.  \textbf{Weakener}

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.  
They are in a library.
Our goal: To generate a rationale that explains why a human would find more likely after learning about a strengthener, and less likely after learning about a weakener.
Our goal: To generate a rationale that explains why a human would find more likely after learning about a strengthener, and less likely after learning about a weaker.
Our goal: To generate a *rationale* that explains why a human would find *more likely* after learning about a *strengthener*, and *less likely* after learning about a *weakenener*.

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting. **Strengthener** They are in a conference room.

A conference room is where people have meetings at work.

You must be quiet in the library, while work meetings involve talking.
Rationale Generation

Motivation

Explainable NLI

Reasoning with Implicit Commonsense Knowledge
Distant Supervision

Why not full supervision?

1. Prior works collected human explanations, which is costly to obtain.
2. Shown to generate very task-specific rationales that doesn’t generalize to other tasks.

Abstract
Deep learning models perform poorly on tasks that require commonsense reasoning, which often necessitates some form of world-knowledge or reasoning over information not immediately present in the input. We collect human explanations for commonsense reasoning in the form of natural language sequences and highlighted annotations in a new dataset called Common Sense Explanations (CoSE). We use CoSE to train language models to automatically generate explanations that can be used during training and inference in a novel Commonsense Auto-Generated Expla-

e-SNLI: Natural Language Inference with Natural Language Explanations

Oana-Maria Camburu1 Tim Rocktäschel2 Thomas Lukasiewicz1,3 Phil Blunsom1,4
1Department of Computer Science, University of Oxford
2Department of Computer Science, University College London
3Alan Turing Institute, London, UK
4DeepMind, London, UK

| Premise: An adult dressed in black holds a stick. | Hypothesis: An adult is walking away, empty-handed. | Label: contradiction |
| Explanation: Holds a stick implies using hands so it is not empty-handed. |

| Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her. | Hypothesis: A young girl is playing with her daughter in a swing. | Label: neutral |
| Explanation: Child does not imply daughter and woman does not imply mother. |

| Premise: A man is in an orange vest leant on a pickup truck. | Hypothesis: A man is touching a truck. | Label: entailment |
| Explanation: Man leans on a pickup truck implies that he is touching it. |

Figure 1: Examples from e-SNLI. Annotators were given the premise, hypothesis, and label. They highlighted the words that they considered essential for the label and provided the explanations.
Overall Framework

Collecting Rationales

\[ \langle P, H, U, T \rangle \rightarrow LM \rightarrow \{R_1, \ldots, R_n\} \]

The definition of \( w_u \) is...
The relationship between \( w_u \) and \( w_h \)...
Before \( H, \ldots \)
After \( P+U, \ldots \)
\( w_h \) implies that...
Overall Framework

Collecting Rationales

<\(P, H, U, T\) \rightarrow LM \rightarrow e-SNLI

Filtering

\{R_1, \ldots, R_n\} \rightarrow e-SNLI Classifier

Top k

\{R_1, \ldots, R_k\}

The definition of \(w_u\) is...
The relationship between \(w_u\) and \(w_h\)...Before \(H\), ...
After \(P + U\), ...
\(w_h\) implies that...
The definition of $w_u$ is...
The relationship between $w_u$ and $w_h$...
Before $H$, ...
After $P+U$, ...
$w_h$ implies that...
Distant Supervision

Vanilla LM

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting. They are in a library.

Step 1:
Recognize salient content words using the attention weights of [CLS] token in defeasible inference classifier.
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting. They are in a library.

Step 1:
Recognize salient content words using the attention weights of [CLS] token in defeasible inference classifier.
Step 1: Recognize salient content words using the attention weights of [CLS] token in defeasible inference classifier.

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting. They are in a library.

Extract top 20% spans w.r.t scores
Distant Supervision

Vanilla LM

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting. They are in a library.

Step 1:
Recognize salient content words using the attention weights of [CLS] token in defeasible inference classifier.

[CLS] 0.4 0.5 0.73 … 0.1 0.67 0.84 0.91

[H] They have a work meeting. [W] They are in a library.

Extract top 20% spans w.r.t scores
Distant Supervision

Vanilla LM

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.  They are in a library.

Step 1: Recognize salient content words using the attention weights of [CLS] token in defeasible inference classifier.

<table>
<thead>
<tr>
<th>[CLS]</th>
<th>0.4</th>
<th>0.5</th>
<th>0.73</th>
<th>...</th>
<th>0.1</th>
<th>0.67</th>
<th>0.84</th>
<th>0.91</th>
</tr>
</thead>
</table>

Extract top 20% spans w.r.t scores

[H] They have a work meeting.  [W] They are in a library.

Post-process: We keep only noun/verb phrase and its head for each span.
Distant Supervision

Vanilla LM

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting. They are in a library.

Step 2: Prompt a pre-trained LM with following context:
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting. They are in a library.

Step 2: Prompt a pre-trained LM with following context:

GPT-2
Distant Supervision

Vanilla LM

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting. They are in a library.

Step 2:
Prompt a pre-trained LM with following context:

GPT-2

P + H + The definition/purpose of $s_{w/u}$ is... The definition of library is that it is a place where people can find books.
**Distant Supervision**

**Vanilla LM**

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting. They are in a library.

---

**Step 2:**
Prompt a pre-trained LM with following context:

GPT-2

\[ P + H \]

+ The definition/purpose of \( S_{w/u} \) is...

\[ P + U + H \]

+ The relationship between \( S_w \) and \( S_u \) is...

---

The definition of **library** is that it is a place where people can find books.

The relationship between **work meeting** and **library** is that you can’t have a meeting in the library.
Distant Supervision

KG-enhanced LM

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting. They are in a library.

<glass of milk, UsedFor, drinking>
Distant Supervision

KG-enhanced LM

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.

They are in a library.

<glass of milk, UsedFor, drinking> → A glass of milk is used for drinking.
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.  

They are in a library.

<glass of milk, UsedFor, drinking>  →  A glass of milk is used for drinking.  →
Distant Supervision

KG-enhanced LM

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting. They are in a library.

The definition/purpose of $s_{h/u}$ is...

The relationship between $s_h$ and $s_u$ is...
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting. They are in a library.

Postcondition dimensions: xWant / xEffect / xReact / xAttr / oWant / oEffect / oReact

Precondition dimensions: xNeed / xIntent / xAttr
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting. They are in a library.

What are postconditions of $u$?

As a result, they want to read.
What are preconditions of $P$?

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting. They are in a library.

What are postconditions of $U$?

As a result, they want to read.

Before, they needed to have a job.

What are preconditions of $H$?
Distant Supervision

NLI-derived

Cardinals lost last night.
The Saint Louis Cardinals always win.

Contradiction explanation: you can’t lose if you always win.

Step 1:
Pre-train WT5 based model on e-SNLI:

Explain nli premise: Cardinals lost last night.
hypothesis: The Saint Louis Cardinals always win.
Distant Supervision

NLI-derived

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.

They are in a library.

Step 2:
Generate Rationale for δ-NLI:

Pre-trained WT5

e-SNLI
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.

They are in a library.

Contradiction explanation:
Being in the library implies being quiet while having a work meeting implies talking.
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting. They are in a library.

Pre-trained model for e-SNLI Using Salient spans

[premise] $w^1_u$ # $w^2_u$ ... [hypothesis] $w^1_h$ # $w^1_h$...

Contradiction

Being in the library implies being quiet.

Having a work meeting implies talking.
Distant Supervision
Filtering Rationales

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting. They are in a library.

List of candidate rationales:

- As a result, they want to read.
- Before, they needed to have a job.
- The definition of library is that it is a place where people can find books.
- The relationship between work meeting and library is that you can't have a meeting in the library.
- Being in the library implies being quiet. Having a work meeting implies talking.

[...]

46
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.  They are in a library.

List of candidate rationales:

As a result, they want to read.
Before, they needed to have a job.
The definition of library is that it is a place where people can find books.
The relationship between work meeting and library is that you can’t have a meeting in the library.
Being in the library implies being quiet.
Having a work meeting implies talking.

[...]
Distant Supervision
Filtering Rationales

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.  They are in a library.

List of candidate rationales:

As a result, they want to read.
Before, they needed to have a job.
The definition of library is that it is a place where people can find books.
The relationship between work meeting and library is that you can't have a meeting in the library.
Being in the library implies being quiet.
Having a work meeting implies talking.

[...]

e-δ-NLI dataset
Distant Supervision

Example rationales

Inputs

P: A person wearing red and white climbs a foggy mountain.
U: The person is attached to a rope going up the side of the mountain.
H: A person is rock climbing.

Extracted Rationale

• 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.

Inputs

P: The brown dog catches a ball in the air.
U: The ball skips into the bushes.
H: The dog plays with the ball outside.

Extracted Rationale

• Catching a ball in the air implies that the dog plays with the ball.
• Bushes are outside.
Training

Post-hoc Rationalization

Encoder

[p] A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

[hp] They have a work meeting.

[u] <Weakener> They are in a library.

Decoder

Being in the library implies being quiet.

Having a work meeting implies talking.
Training

Post-hoc Rationalization

Encoder

[prefm] A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

[hpthesis] They have a work meeting.

[update_type] <Weakener> [update] They are in a library.

Decoder

Being in the library implies being quiet.
Having a work meeting implies talking.

Joint Prediction and Rationalization

Encoder

[prefm] A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

[hpthesis] They have a work meeting.

[update_type] <Weakener>

Decoder

They are in a library.
Being in the library implies being quiet.
Having a work meeting implies talking.
Training

Objectives
Training

Objectives

Underlying LM

\{GPT-2, Bart\}
## Training Objectives

<table>
<thead>
<tr>
<th>Underlying LM</th>
<th>Training Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>{GPT-2, Bart}</td>
<td>{rationale, multi, update+rationale update-type+rationale}</td>
</tr>
</tbody>
</table>

Eight different setups
Underlying LM | Training Objective
---|---
{GPT-2, Bart} × {rationale, multi, update+rationale update-type+rationale}

Rationale: \( P(R \mid P, H, T, U) \)

Eight different setups
Training
Objectives

Underlying LM
\{GPT-2, Bart\} \times \{\text{rationale, multi, update+rationale, update-type+rationale}\}

<table>
<thead>
<tr>
<th>Post-hoc Rat.</th>
<th>Training Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rationale:</strong></td>
<td>(P(R</td>
</tr>
<tr>
<td><strong>Multi-task:</strong></td>
<td>(P(R</td>
</tr>
</tbody>
</table>

Eight different setups
Training

Objectives

Underlying LM \{GPT-2, Bart\} x \{rationale, multi, update+rationale update-type+rationale\}

**Post-hoc Rat.**

- **Rationale:** \(P(R | P, H, T, U)\)
- **Multi-task:**
  - \(P(R | P, H, T, U)\)
  - \(P(U | P, H, T, R)\)
  - \(P(T | P, H, U, R)\)

**Joint Pred. + Rat.**

- **Update+Rationale:** \(P(R, U | P, H, T)\)
- **Update-type+Rationale:** \(P(R, T | P, H, U)\)

Eight different setups
Results

Automatic Metrics*
Results

Automatic Metrics*

**BLEU-4**

- **GPT2-XL**: 33
- **BART-Large**: 13.15

**ROUGE-L**

- **GPT2-XL**: 33.91
- **BART-Large**: 22.48

* on the distant supervision
Results

Automatic Metrics*

**BLEU-4**

- GPT2-XL: 33, Multi: 33.58
- BART-Large: 13.15, type+Rationale: 33.58

**ROUGE-L**

- GPT2-XL: 33.91, Multi: 34.5
- BART-Large: 22.48, Type+Rationale: 24.03

* on the distant supervision
Results

Automatic Metrics*

**BLEU-4**

- Rationale: 33.15
- Multi: 17.38
- Update+Rationale: 23.93
- type+Rationale: 25.24

**ROUGE-L**

- Rationale: 33.91
- Multi: 22.48
- Update+Rationale: 31.71
- Type+Rationale: 30.83

* on the distant supervision
Results

Automatic Metrics*

- **GPT2 > Bart**
- **Multi performs best**
- **Post-hoc is performing better showing joint is harder**

* on the distant supervision
On 200 sampled instances:
1. Grammaticality
2. **Relevant** to instance
3. Factually **correct** or likely true
4. **Explanatory**

**Premise:** A guy riding a motorcycle near junk cars.

**Hypothesis:** The man is test driving a motorcycle to decide whether or not he will buy it.

**Weakener:** The man is wearing a bright outfit and there are hundreds of people cheering for him.

**Rationale 1:** The man is wearing a bright outfit and there are hundreds of people cheering for him. The definition of "man" is defined as male human.

| The rationale is completely gibberish, I can't understand it at all. |
| The rationale is not perfectly grammatical, but I can understand it. |
| The rationale is grammatical. |

| The rationale is on topic with respect to the premise and hypothesis. |
| The rationale is factually correct or likely true. |
| The rationale may explain why the context weakens the hypothesis. |
## Evaluation

### Human Eval.

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Correct</th>
<th>Explains</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPT-2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rationale</td>
<td>82%</td>
<td>30%</td>
<td>47%</td>
</tr>
<tr>
<td>Multi</td>
<td>0.5%</td>
<td>15.5%</td>
<td>30.5%</td>
</tr>
<tr>
<td>Update+ Rationale</td>
<td>1%</td>
<td>15.5%</td>
<td>30.5%</td>
</tr>
<tr>
<td>Type+ Rationale</td>
<td>8.5%</td>
<td>27%</td>
<td>36.5%</td>
</tr>
<tr>
<td><strong>BART</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rationale</td>
<td>8.5%</td>
<td>33.5%</td>
<td>62.5%</td>
</tr>
<tr>
<td>Multi</td>
<td>0%</td>
<td>34%</td>
<td>36%</td>
</tr>
<tr>
<td>Update+ Rationale</td>
<td>26%</td>
<td>34%</td>
<td>36.5%</td>
</tr>
<tr>
<td>Type+ Rationale</td>
<td>26%</td>
<td>35.5%</td>
<td></td>
</tr>
</tbody>
</table>

### Joint generation is more challenging

**Bart >> GPT2**

**Multi-tasking doesn't help**

(Almost everything is grammatical 83%-99%)
Evaluation

Weakener vs. Strengthener in being explanatory?
Evaluation

Weakener vs. Strengthener in being explanatory?

Most rationales are relevant, about half are correct, and between 1/3 and 1/2 explain the update type.
Most rationales are relevant, about half are correct, and between 1/3 and 1/2 explain the update type.

- It is easier to generate rationales for strengtheners!
Analysis

Quality of Generated Rationales Evaluated as Explanatory by Humans
Analysis
Quality of Generated Rationales Evaluated as Explanatory by Humans

Strengthener

<table>
<thead>
<tr>
<th>Pattern</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>[S] ([H]) implies (that) [H] ([S])</td>
<td>64.9</td>
</tr>
<tr>
<td>[S] ([H]) is a rephrasing of [H] ([S])</td>
<td>14.9</td>
</tr>
<tr>
<td>[H] ([S]) because [S] ([H])</td>
<td>12.8</td>
</tr>
<tr>
<td>[S] means [H]</td>
<td>2.1</td>
</tr>
<tr>
<td>[S] is [H]</td>
<td>1.1</td>
</tr>
<tr>
<td>[S] is the same as [H]</td>
<td>1.1</td>
</tr>
<tr>
<td>Other</td>
<td>3.9</td>
</tr>
</tbody>
</table>
Analysis

Quality of Generated Rationales Evaluated as Explanatory by Humans

Strengthener

<table>
<thead>
<tr>
<th>Pattern</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>([S] ([H])) implies (that) ([H] ([S]))</td>
<td>64.9</td>
</tr>
<tr>
<td>([S] ([H])) is a rephrasing of ([H] ([S]))</td>
<td>14.9</td>
</tr>
<tr>
<td>([H] ([S])) because ([S] ([H]))</td>
<td>12.8</td>
</tr>
<tr>
<td>([S]) means ([H])</td>
<td>2.1</td>
</tr>
<tr>
<td>([S]) is ([H])</td>
<td>1.1</td>
</tr>
<tr>
<td>([S]) is the same as ([H])</td>
<td>1.1</td>
</tr>
<tr>
<td>Other</td>
<td>3.9</td>
</tr>
</tbody>
</table>

Weakener

<table>
<thead>
<tr>
<th>Pattern</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Something cannot be ([W]) and ([H]) at the same time</td>
<td>33.3</td>
</tr>
<tr>
<td>Something cannot be ([W]) ([H]) if it is ([H]) ([W])</td>
<td>31.8</td>
</tr>
<tr>
<td>([W]) is not the same as ([H])</td>
<td>13.6</td>
</tr>
<tr>
<td>Something is either ([W]) or ([H])</td>
<td>10.6</td>
</tr>
<tr>
<td>([W]) is not ([H])</td>
<td>6.1</td>
</tr>
<tr>
<td>Other</td>
<td>4.6</td>
</tr>
</tbody>
</table>

Almost all of them fit into one of several patterns that are trivial to generate given the update type.
Analysis

Ablation Studies: 1. filtering step, 2. NLI-derived only rationales

- Both ablations increase the relevance of rationales while hurting their factual correctness and producing less explanatory weakener rationales.
Analysis

Ablation Studies: 1. filtering step, 2. NLI-derived only rationales

Both ablations increase the relevance of rationales while hurting their factual correctness and producing less explanatory weakener rationales.

Hypothesis: 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.
Takeaways
Takeaways

🤖 Multi-tasking did not improve the rationale generation performance.
Takeaways

🤖 Multi-tasking did not improve the rationale generation performance.

🤖 Generating rationales given the label (post-hoc) is less challenging
   🤖 Many of them considered as explanatory by humans
   🤖 Cheat → learnt the trivial mapping: “[update] implies that [hypothesis] is less/more likely”
Takeaways

🤖 Multi-tasking did not improve the rationale generation performance.

🤖 Generating rationales given the label (post-hoc) is less challenging
   🤖 Many of them considered as explanatory by humans
   🤖 Cheat → learnt the trivial mapping: “[update] implies that [hypothesis] is less/more likely”

🤖 Jointly predicting update & generating rationale: extremely challenging.
Takeaways

🤖 Multi-tasking did not improve the rationale generation performance.

🤖 Generating rationales given the label (post-hoc) is less challenging
  Many of them considered as explanatory by humans
  Cheat —> learnt the trivial mapping: “[update] implies that [hypothesis] is less/more likely”

🤖 Jointly predicting update & generating rationale: extremely challenging.

Future directions:
- Focus on jointly predicting a label and generating rationale —> less trivial and more faithful.