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

why?

what?

how?

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!

A Logic for Default Reasoning. Raymond Reiter. Artificial Intelligence, 1980.

Defeasible Inference

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

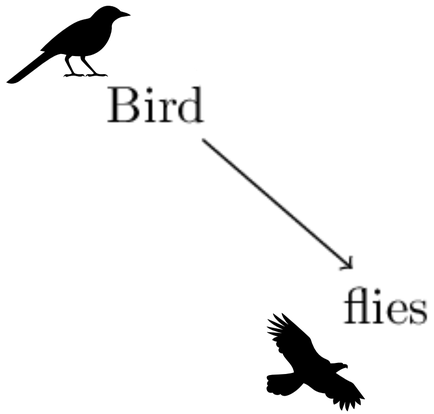
Bird

P: Tweety is a bird.

A Logic for Default Reasoning. Raymond Reiter. Artificial Intelligence, 1980.

Defeasible Inference

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



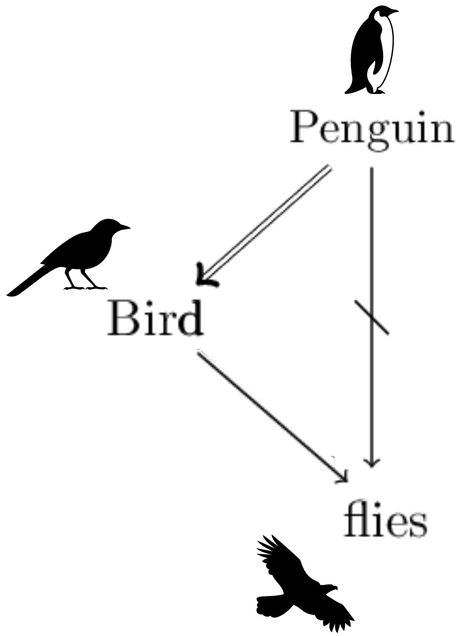
P: Tweety is a bird.

H: Tweety flies.

A Logic for Default Reasoning. Raymond Reiter. Artificial Intelligence, 1980.

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.

H: Tweety flies.

U: Tweety is a penguin.

A Logic for Default Reasoning. Raymond Reiter. Artificial Intelligence, 1980.

Defeasible Inference

Given a premise 𝔓 and hypothesis 𝔥, an update 𝔲 is called:

P: Tweety is a bird. 

H: Tweety flies. 

Thinking Like a Skeptic: Defeasible Inference in Natural Language.
Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith and Yejin Choi. Findings of EMNLP, 2020.

Defeasible Inference

Given a premise 𝔓 and hypothesis 𝔥, an update 𝔲 is called:

weakener —> if a human would most likely find 𝔥 *less likely to be true* after learning 𝔲;

P: Tweety is a bird. 

H: Tweety flies. 

Weakener: Tweety is a penguin. 

Thinking Like a Skeptic: Defeasible Inference in Natural Language.
Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith and Yejin Choi. Findings of EMNLP, 2020.

Defeasible Inference

Given a premise P and hypothesis H, an update U is called:

weakener —> if a human would most likely find H *less likely to be true* after learning U;

strengthener —> if they would find H *more likely to be true*

P: Tweety is a bird. 

H: Tweety flies. 

Weakener: Tweety is a penguin. 

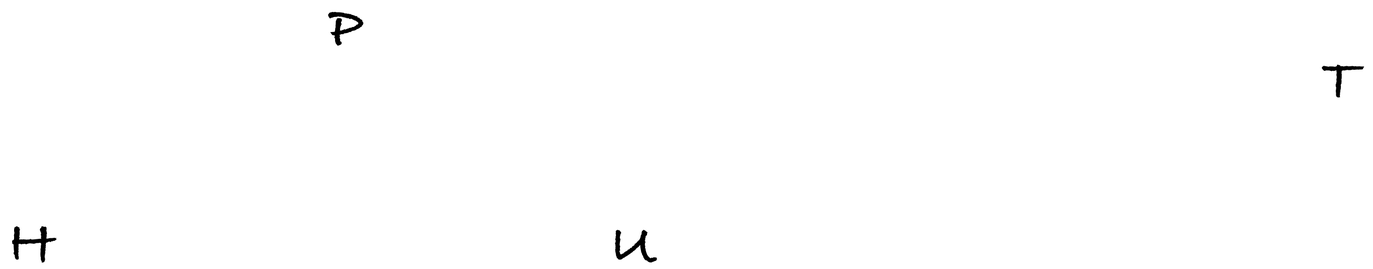
Strengthener: Tweety is on a tree. 

Thinking Like a Skeptic: Defeasible Inference in Natural Language.
Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith and Yejin Choi. Findings of EMNLP, 2020.

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.

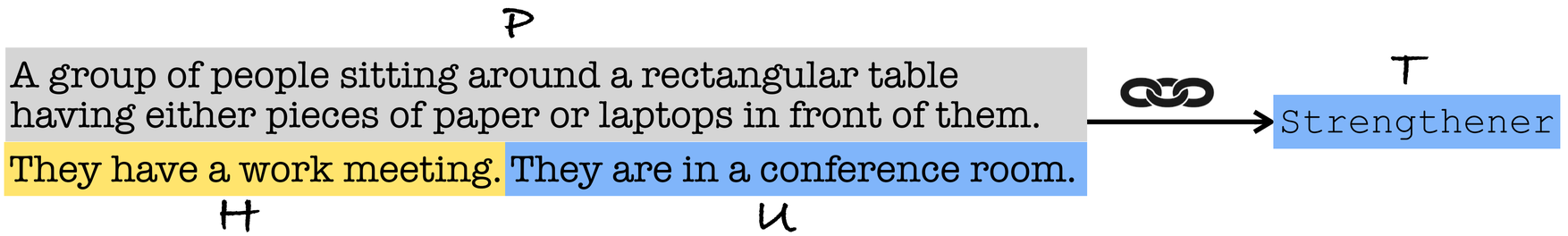


Thinking Like a Skeptic: Defeasible Inference in Natural Language.
Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith and Yejin Choi. Findings of EMNLP, 2020.

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.

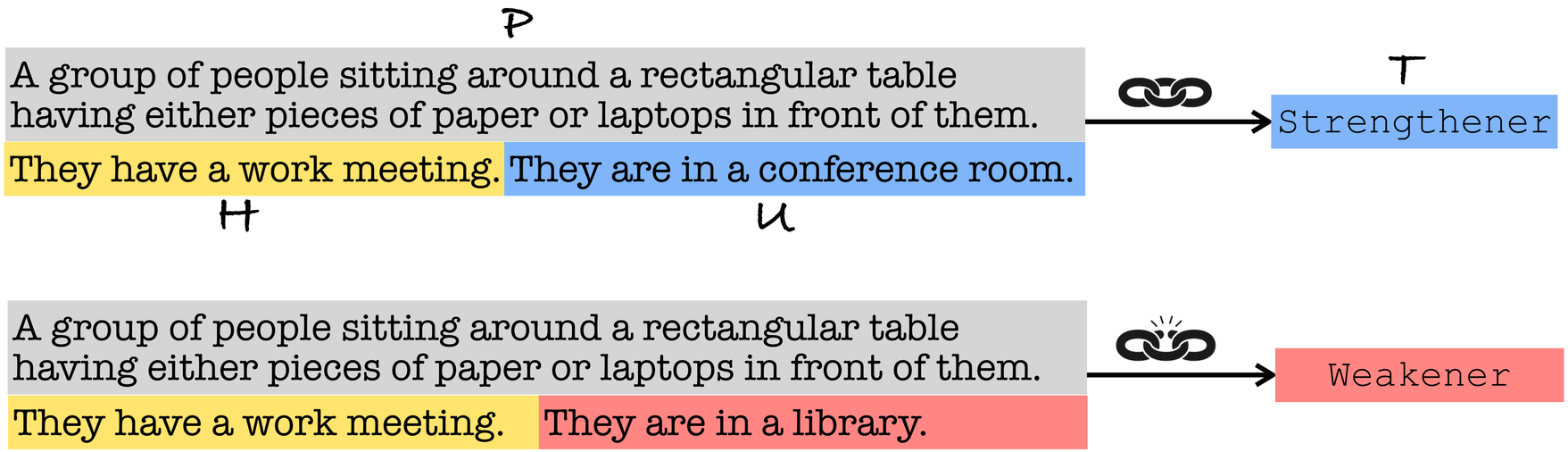


Thinking Like a Skeptic: Defeasible Inference in Natural Language.
Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith and Yejin Choi. Findings of EMNLP, 2020.

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.

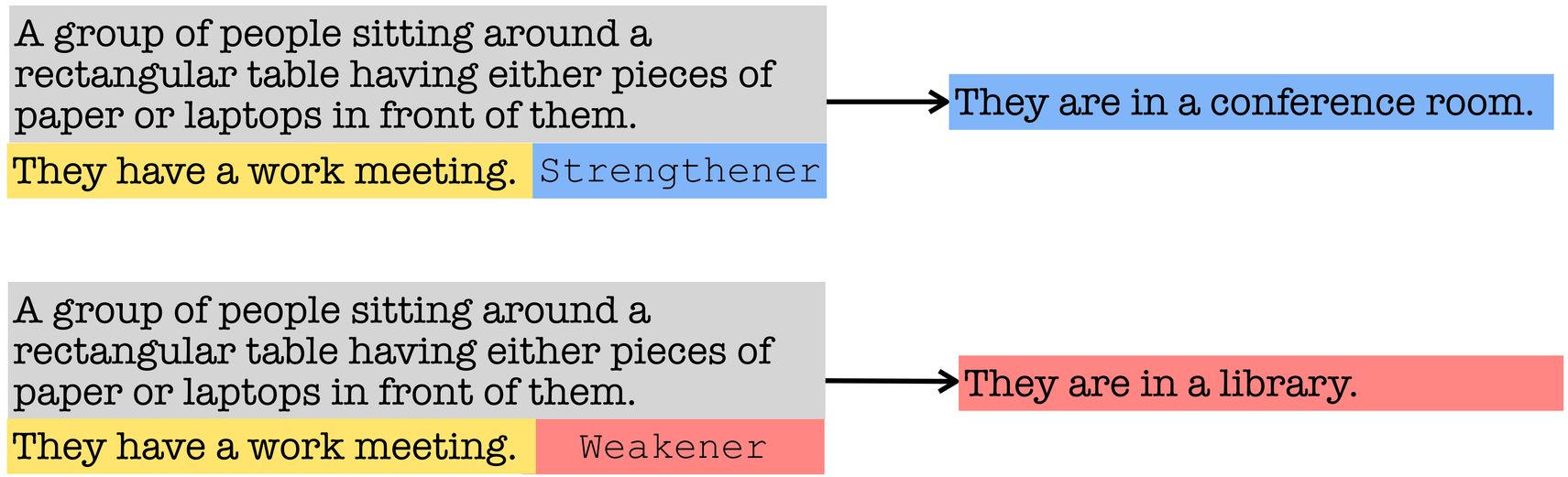


Thinking Like a Skeptic: Defeasible Inference in Natural Language.
Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith and Yejin Choi. Findings of EMNLP, 2020.

Defeasible Inference

Generative Task

Given a ***premise*** P, a ***hypothesis*** H, and a desired update type T (**weakener** or **strengthener**), the goal is to generate an ***update*** U that satisfies the type constraint.



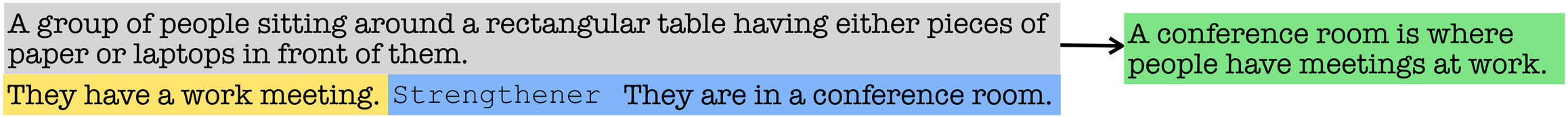
Thinking Like a Skeptic: Defeasible Inference in Natural Language.
Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith and Yejin Choi. Findings of EMNLP, 2020.

Rationale Generation

Our goal: To generate a ***rationale*** r that explains <u>why</u> a human would find h *more likely* after learning about a **strengthener**, and *less likely* after learning about a **weakener**.

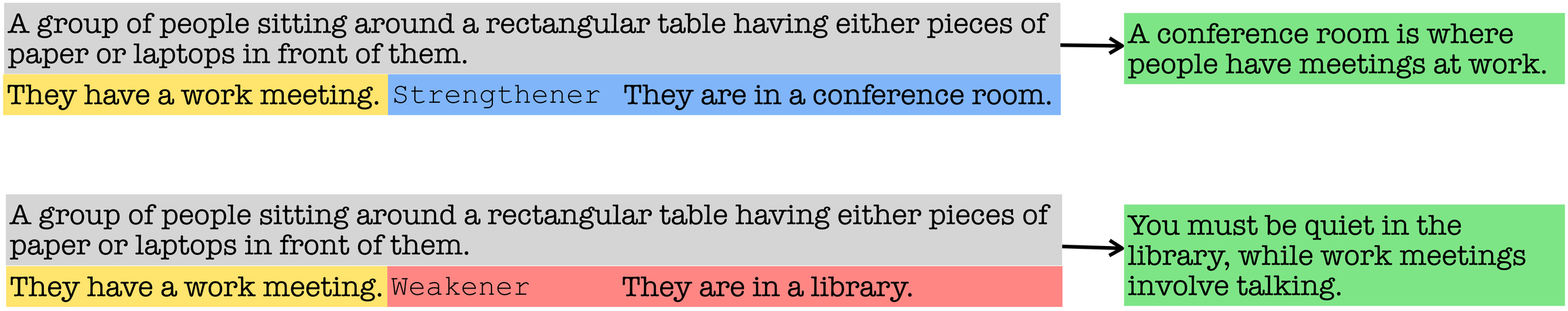
Rationale Generation

Our goal: To generate a ***rationale*** r that explains <u>why</u> a human would find h *more likely* after learning about a **strengthener**, and *less likely* after learning about a **weakener**.



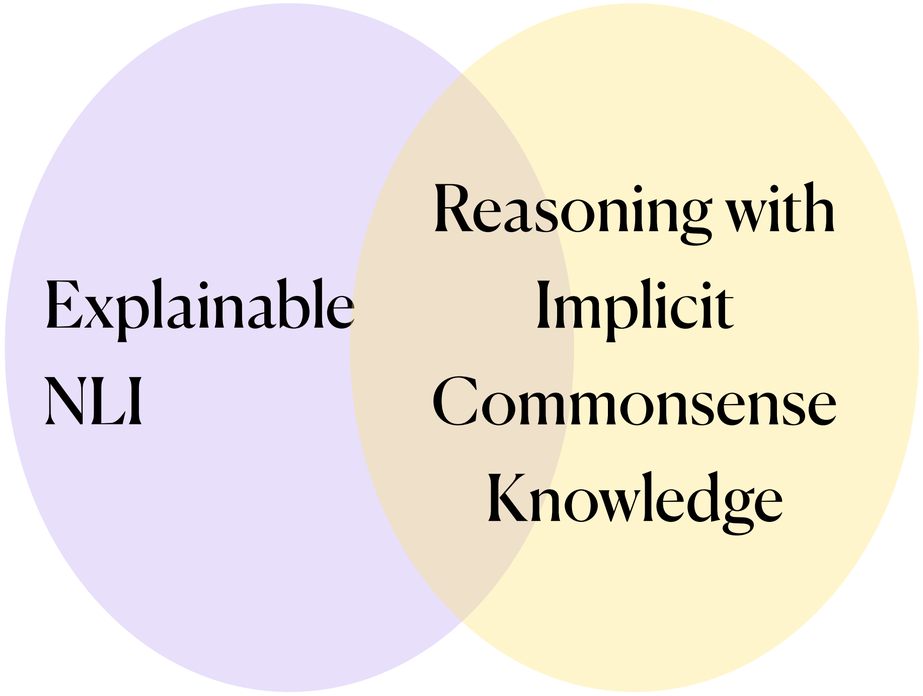
Rationale Generation

Our goal: To generate a ***rationale*** r that explains <u>why</u> a human would find h *more likely* after learning about a **strengthener**, and *less likely* after learning about a **weakener**.



Rationale Generation

Motivation



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.

Explain Yourself!
Leveraging Language Models for Commonsense Reasoning

Nazneen Fatema Rajani **Bryan McCann** **Caiming Xiong** **Richard Socher**
Salesforce Research
Palo Alto, CA, 94301
{nazneen.rajani,bmccann,cxiong,rsocher}@salesforce.com

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 (CoS-E). We use CoS-E to train language models to automatically generate explanations that can be used during training and inference in a novel Commonsense Auto-Generated Expla-

Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.
Question:	After getting drunk people couldn't understand him,it was because of his what?
Choices:	lower standards,**slurred speech**, or falling down
CoS-E:	People who are drunk have difficulty speaking.
Question:	People do what during their time off from work?
Choices:	**take trips**, brow shorter, or become hysterical
CoS-E:	People usually do something relaxing, such as taking trips,when they don't need to work.

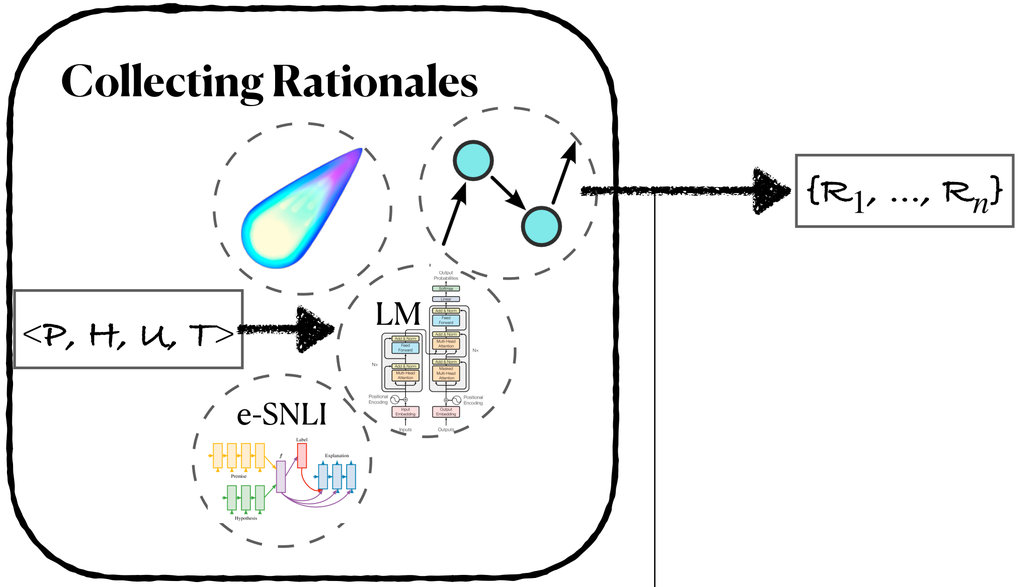
e-SNLI: Natural Language Inference with Natural Language Explanations

Oana-Maria Camburu[1] **Tim Rocktäschel**[2] **Thomas Lukasiewicz**[1,3] **Phil Blunsom**[1,4]
{oana-maria.camburu, thomas.lukasiewicz, phil.blunsom}@cs.ox.ac.uk
t.rocktaschel@ucl.ac.uk
[1]Department of Computer Science, University of Oxford
[2]Department of Computer Science, University College London
[3]Alan Turing Institute, London, UK
[4]DeepMind, London, UK

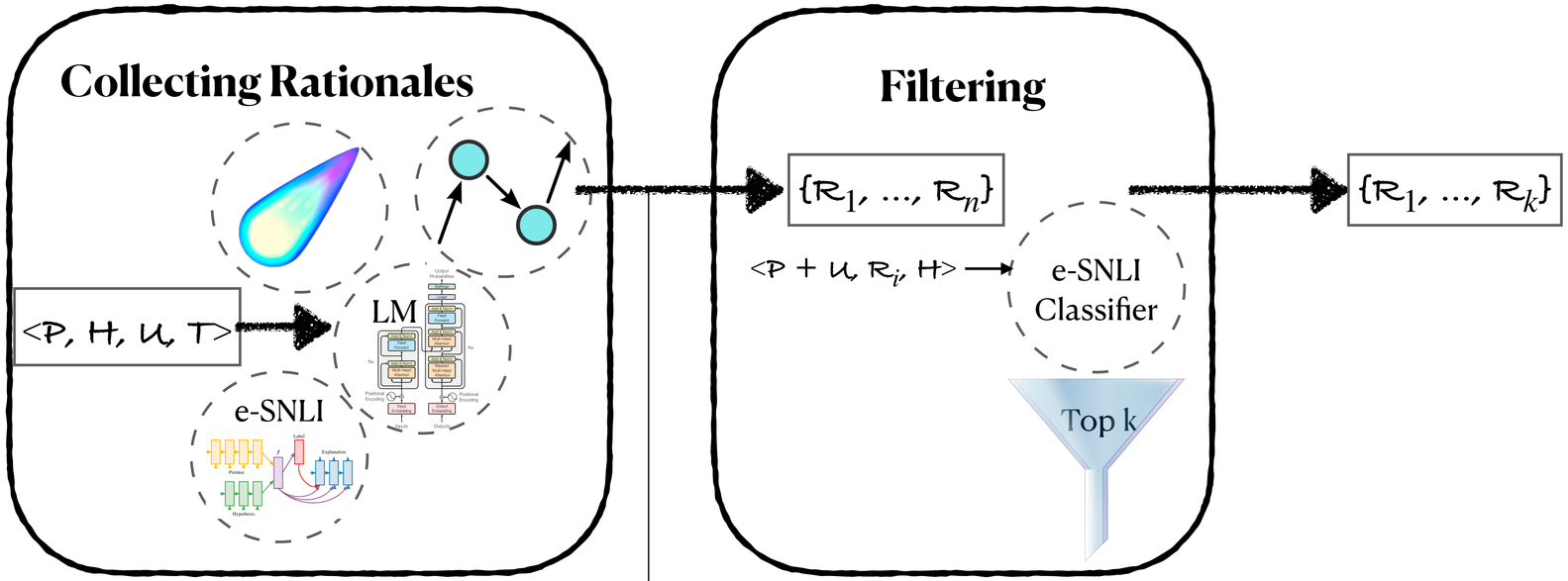
Premise: An adult dressed in black holds a stick. Hypothesis: An adult is walking away, empty-handed. Label: contradiction Explanation: Holds a stick implies using hands so it is not empty-handed.
Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her. Hypothesis: A young mother 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 in an orange vest leans over a pickup truck. Hypothesis: A man is touching a truck. Label: entailment Explanation: Man leans over 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

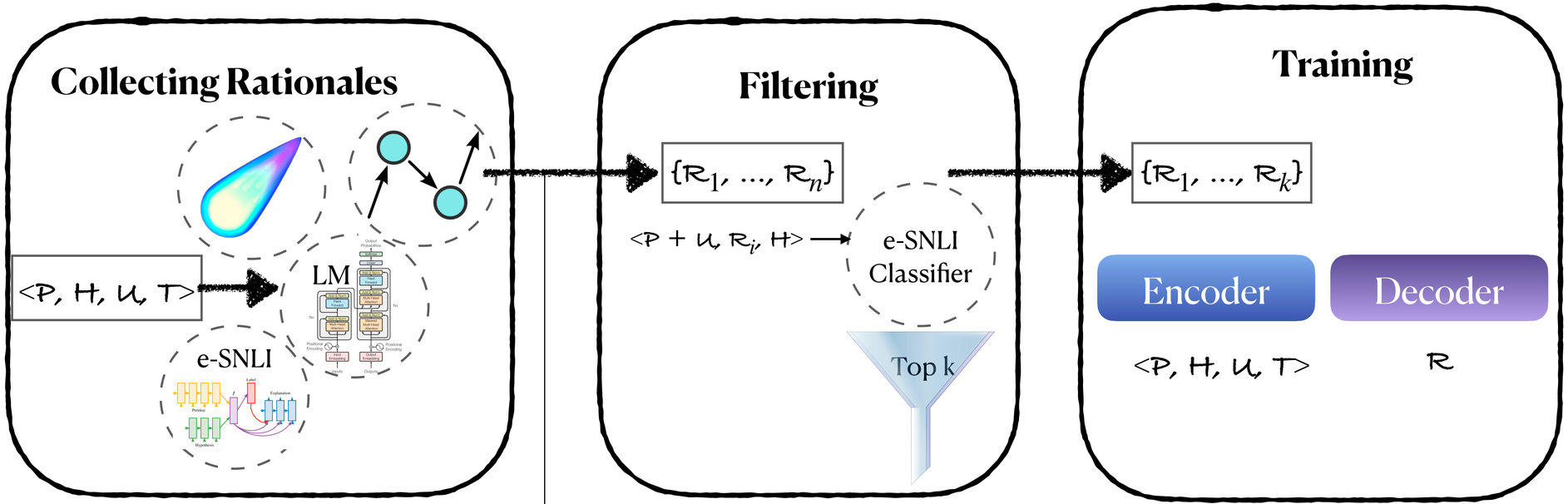


Overall Framework



The definition of w_u is...
The relationship between w_u and w_h...
Before H, ...
After P+U,...
w_h implies that...

Overall Framework



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.

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.

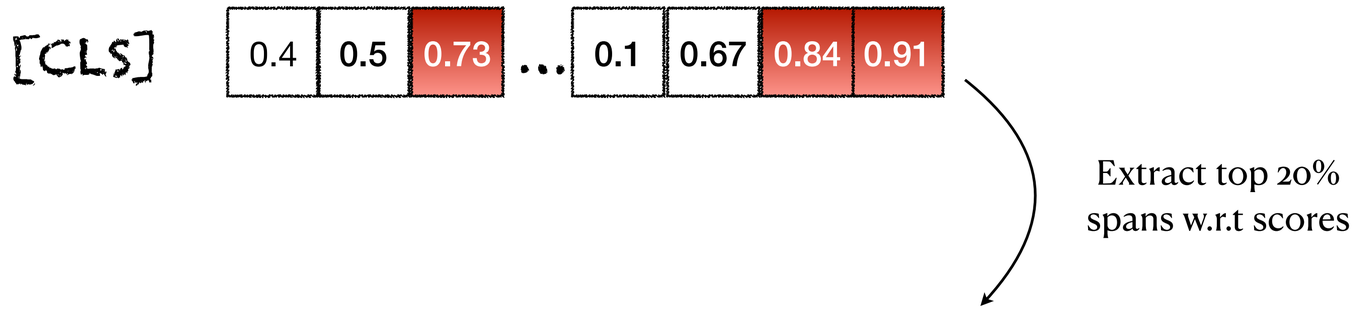
Distant Supervision

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.



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.

Extract top 20% spans w.r.t scores

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

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.

Extract top 20% spans w.r.t scores

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

s_h s_u

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:

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

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:

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

GPT-2

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 $s_{h/u}$ is...

P + U + H + The relationship between s_h 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.

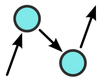
Distant Supervision

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>

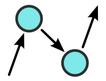
Distant Supervision

KG-enhanced LM

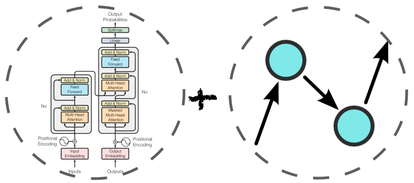
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.

They are in a library.



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



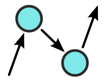
Distant Supervision

KG-enhanced LM

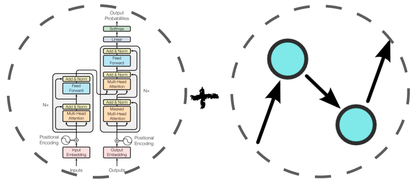
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.

They are in a library.



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

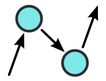


Distant Supervision

KG-enhanced LM

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

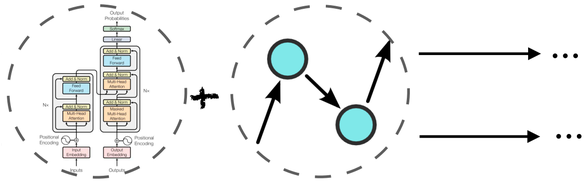
They have a work meeting. | They are in a library.



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

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

The relationship between s_h and s_u is...

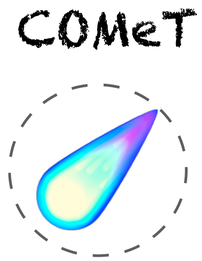


Distant Supervision

COMeT

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

H They have a work meeting. | They are in a library. U



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

Precondition dimensions:
xNeed / xIntent / xAttr

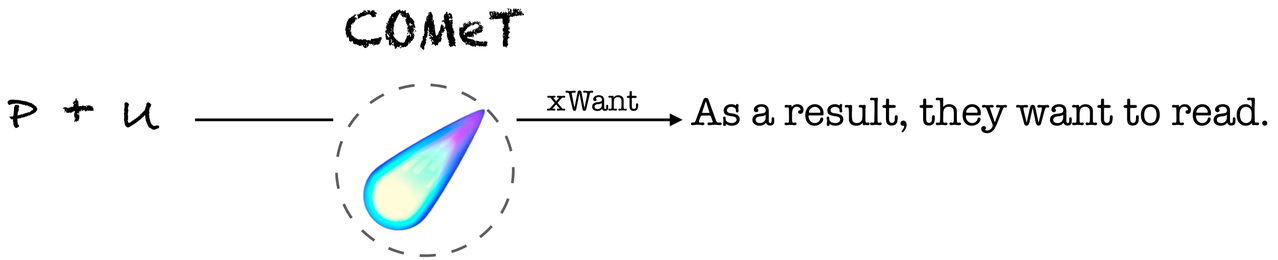
Distant Supervision

COMeT

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

H They have a work meeting. | They are in a library. u

What are postconditions of u?

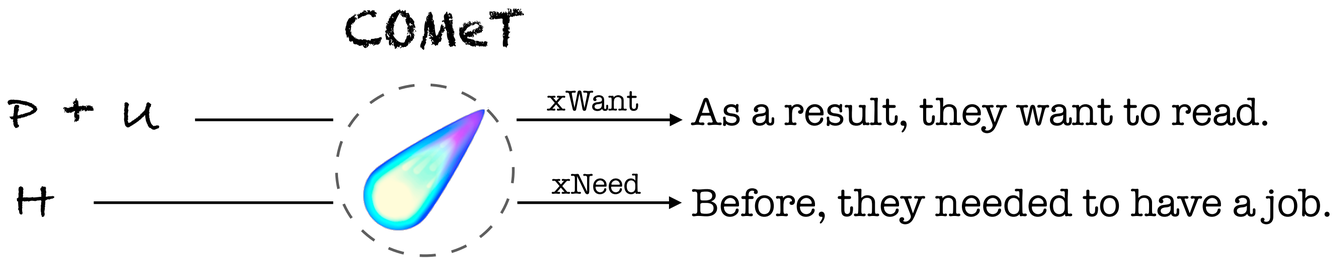


Distant Supervision

COMeT

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

What are postconditions of U?



What are preconditions of H?

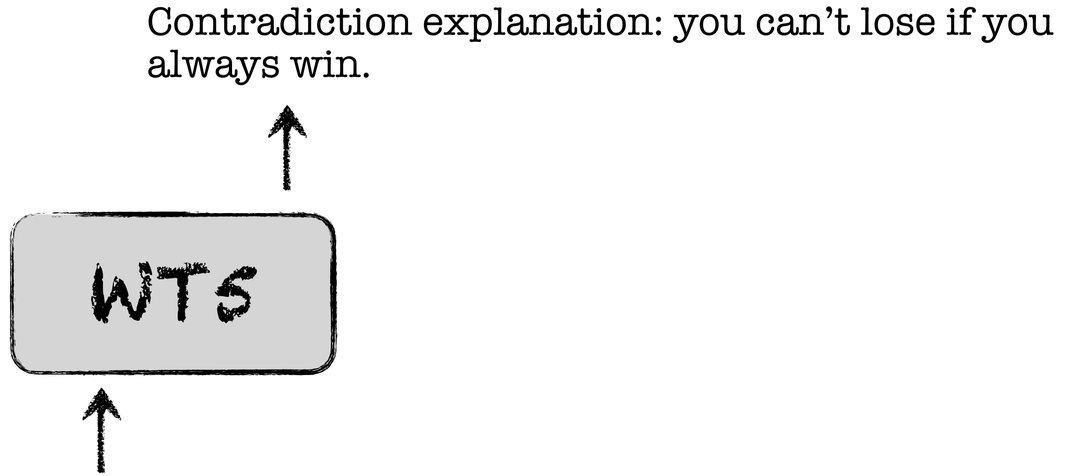
Distant Supervision

NLI-derived

Cardinals lost last night.

The Saint Louis Cardinals always win.

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



WT5?! Training Text-to-Text Models to Explain their Predictions. Narang et al., 2020

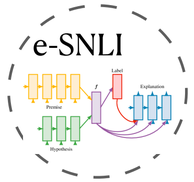
Distant Supervision

NLI-derived

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

Step 2:
Generate Rationale for
δ-NLI:

Pre-trained WTs



Distant Supervision

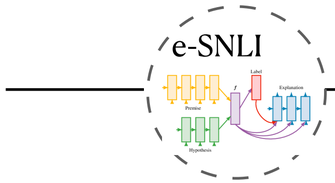
NLI-derived

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.

Pre-trained WTs

P A group of people sitting around a rectangular table [...] or laptops [...].
U They are in a library. <sep>
H They have a work meeting.



Contradiction explanation:
Being in the library implies being quiet while having a work meeting implies talking.

Distant Supervision

NLI-derived w/ Highlights

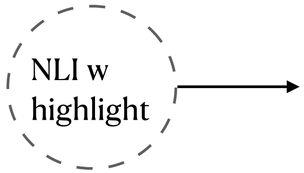
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.

They are in a library.

Pre-trained model for e-SNLI Using Salient spans

[premise] w_u^1# w_u^2 ... [hypothesis] w_h^1# w_h^1...



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.

[...]

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.

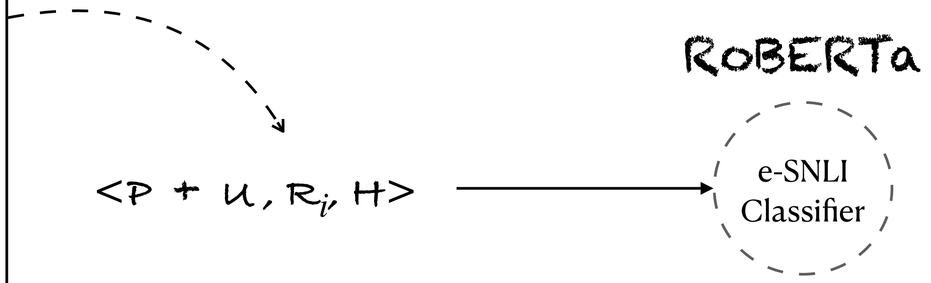
Before, they needed to have a job.

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.

[...]



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.

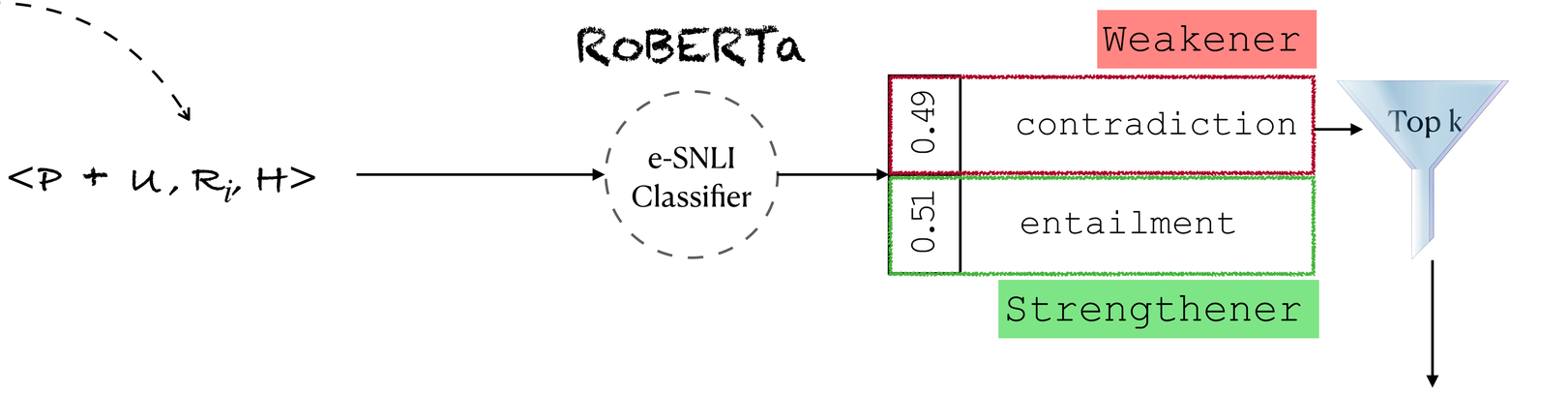
Before, they needed to have a job.

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

KG-enhanced LM

Inputs

$\mathcal{P}$: A person wearing red and white climbs a foggy mountain.
$\mathcal{U}_s$: The person is attached to a rope going up the side of the mountain.
$\mathcal{H}$: 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.

NLI-derived

Inputs

$\mathcal{P}$: The brown dog catches a ball in the air.
$\mathcal{U}_s$: The ball skips into the bushes.
$\mathcal{H}$: 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	Decoder
[premise] A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.	Being in the library implies being quiet. Having a work meeting implies talking.
[hypothesis] They have a work meeting.	
[update_type] `<Weakener>` [update] They are in a library.	

Training

Post-hoc Rationalization

Encoder	Decoder
[premise] A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.	Being in the library implies being quiet. Having a work meeting implies talking.
[hypothesis] They have a work meeting.	
[update_type] <Weakener> [update] They are in a library.	

Joint Prediction and Rationalization

Encoder	Decoder
[premise] 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.
[hypothesis] They have a work meeting.	Being in the library implies being quiet. Having a work meeting implies talking.
[update_type] <Weakener>	

Training

Objectives

Training

Objectives

Underlying LM

{GPT-2, Bart}

Training

Objectives

Underlying LM		Training Objective
{GPT-2, Bart}	X	{rationale, multi, update+rationale update-type+rationale}

Eight different setups

Training

Objectives

Underlying LM | Training Objective

{GPT-2, Bart} X {rationale, multi, update+rationale update-type+rationale}

Post-hoc Rat.

Rationale: P(R| P, H, T, U)

Eight different setups

Training

Objectives

Underlying LM | Training Objective

{GPT-2, Bart} X {rationale, multi, update+rationale update-type+rationale}

Post-hoc Rat.

Rationale:	$P(R \mid P, H, T, U)$
Multi-task:	$P(R \mid P, H, T, U)$ $P(U \mid P, H, T, R)$ $P(T \mid P, H, U, R)$

Eight different setups

Training

Objectives

Underlying LM: {GPT-2, Bart} X Training Objective: {rationale, multi, update+rationale update-type+rationale}

Eight different setups

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: $P(R, T|\ P, H, U)$

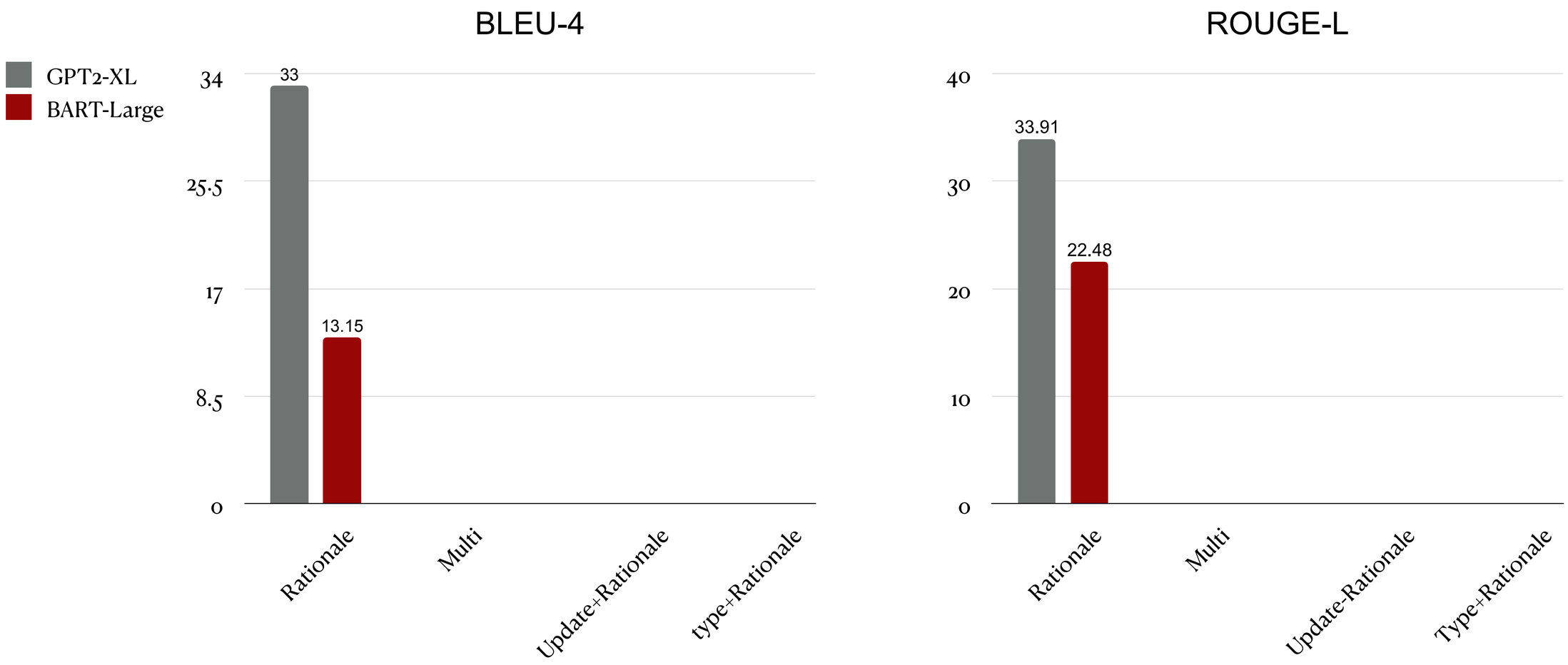
Results

Automatic Metrics*

* on the distant supervision

Results

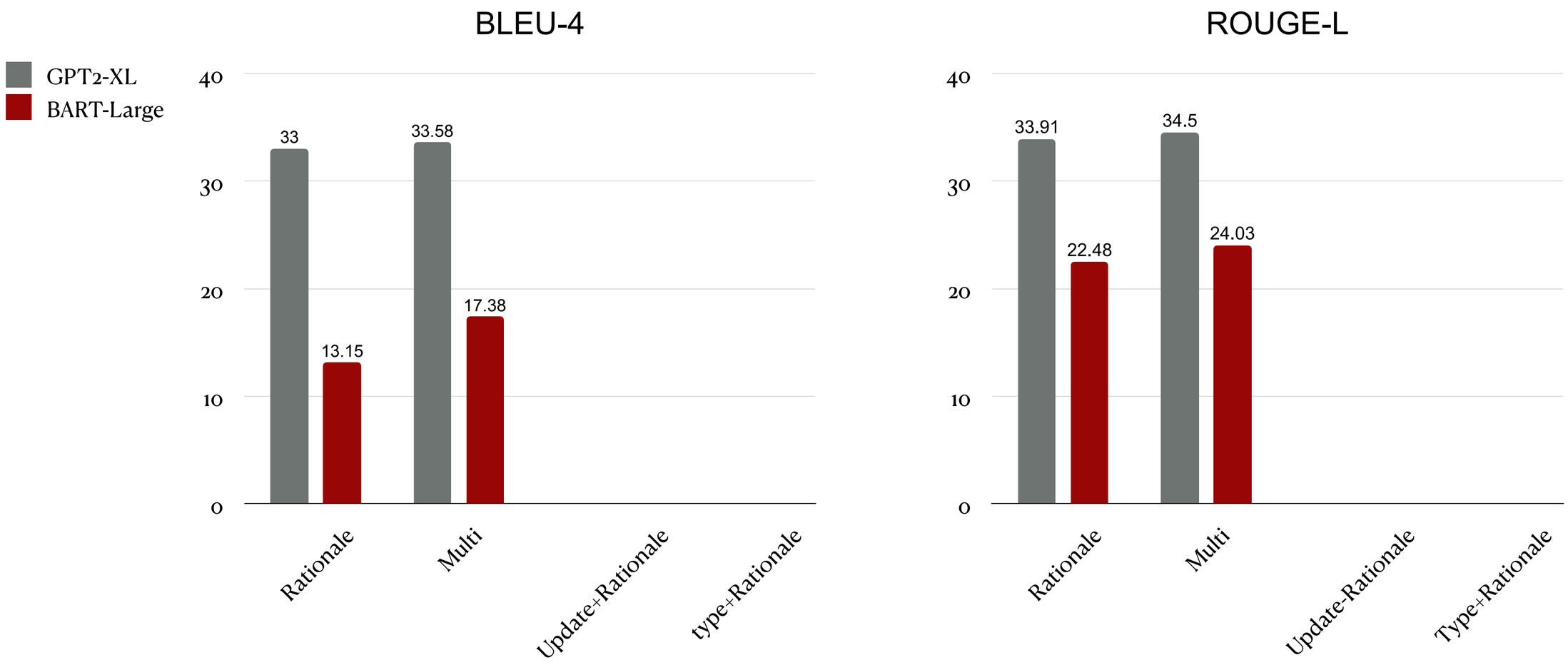
Automatic Metrics*



* on the distant supervision

Results

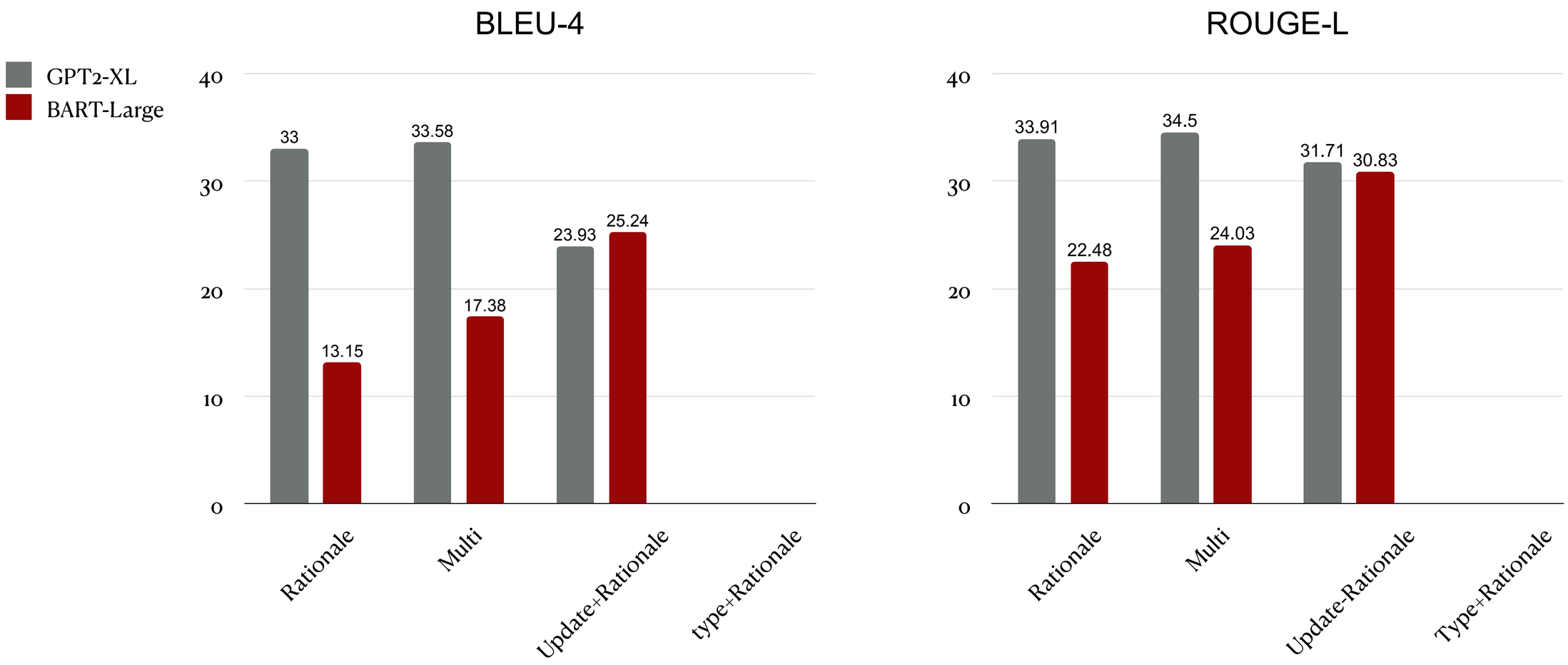
Automatic Metrics*



* on the distant supervision

Results

Automatic Metrics*

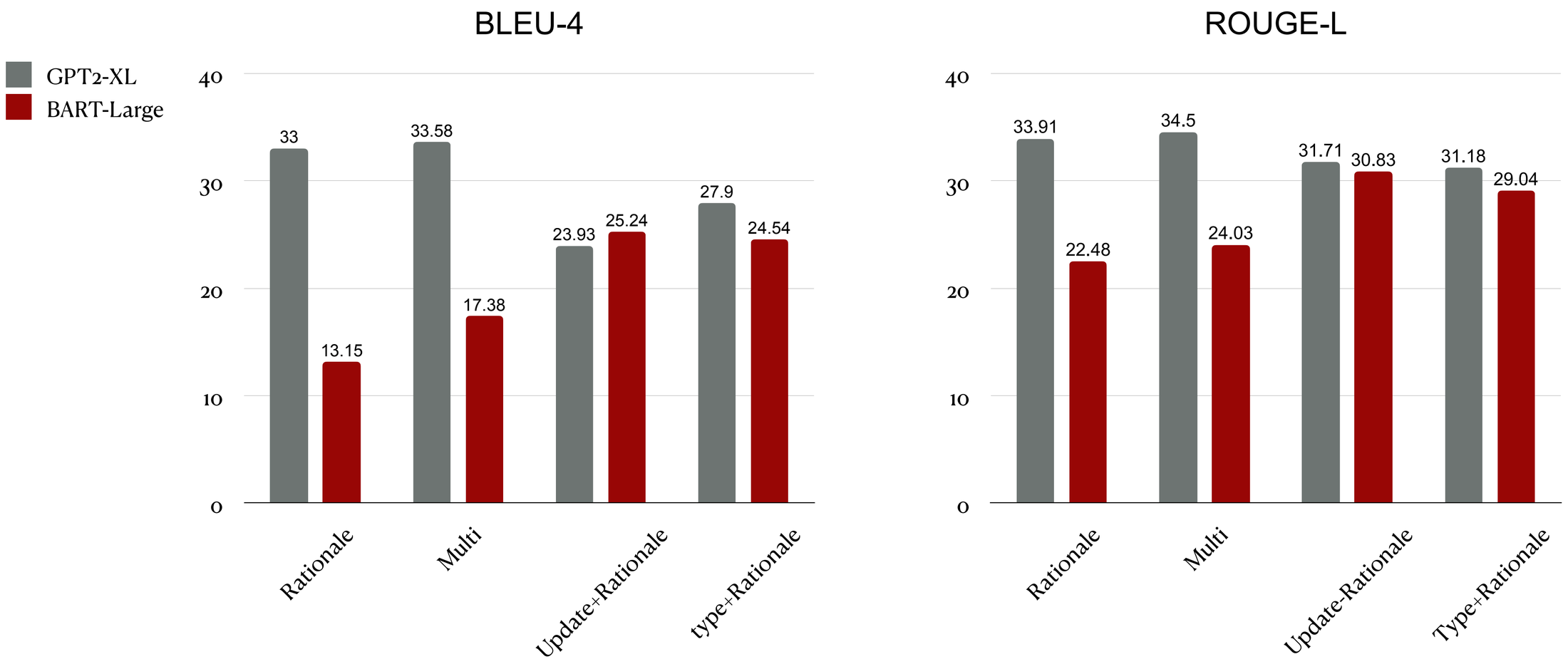


* on the distant supervision

Results

Automatic Metrics*

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



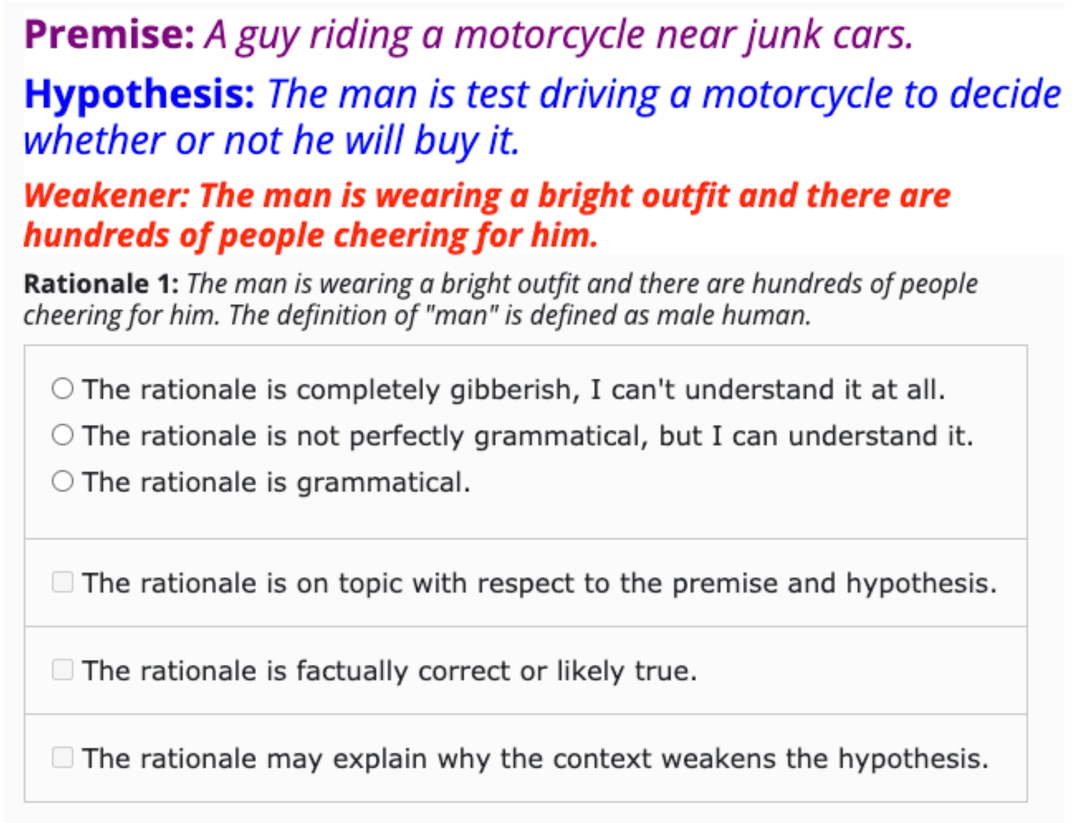
* on the distant supervision

Evaluation

Human Eval.

On 200 sampled instances:

1. Grammaticality
2. **Relevant** to instance
3. Factually **correct** or likely true
4. **Explanatory**



Evaluation

Human Eval.

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

☑ Multi-tasking doesn't help

☑ Joint generation is more challenging

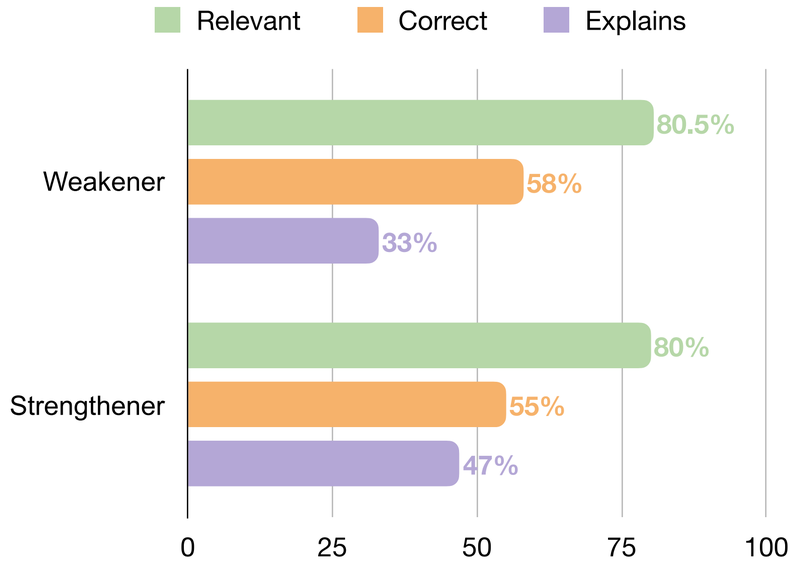
☑ Bart >> GPT2

Evaluation

Weakener vs. Strengthener in being explanatory?

Evaluation

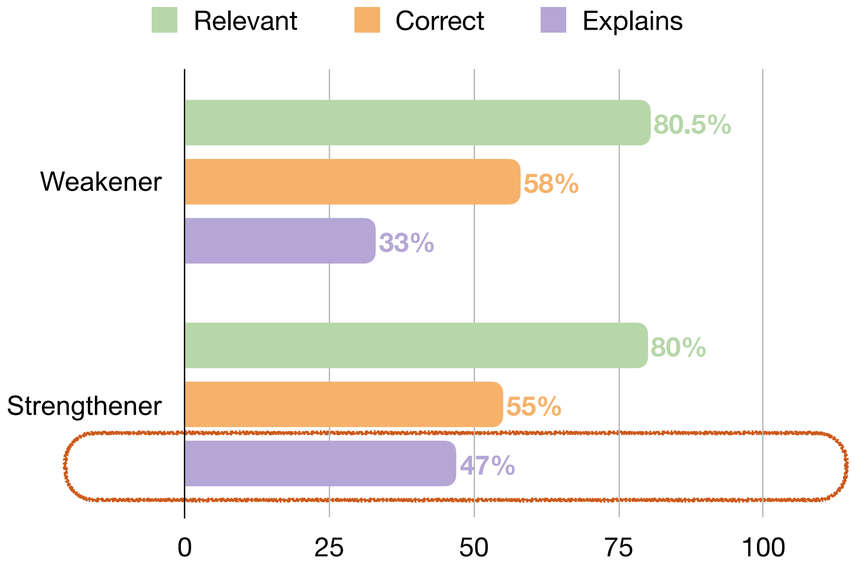
Weakener vs. Strengthener in being explanatory?



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

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

☑ It is easier to generate rationales for strengthener!

Analysis

Quality of Generated Rationales Evaluated as Explanatory by Humans

Analysis

Quality of Generated Rationales Evaluated as Explanatory by Humans

Strengthener

Pattern	%
[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	3.9

Analysis

Quality of Generated Rationales Evaluated as Explanatory by Humans

Strengthener

Pattern	%
[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	3.9

Weakener

Pattern	%
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

- ☑ 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.

Code

fbrahman@ucsc.edu

@faeze_brh