Evaluating the Rationale Understanding for Critical Reasoning in Logical Reading Comprehension

Akira Kawabata
The Asahi Shimbun Company

Saku Sugawara
National Institute of Informatics

Paper

Our Contributions

- We create a dataset to evaluate the understanding of the rationale in logical reasoning.
- Current models struggle to comprehend the rationale for eliminating plausible alternatives.

Motivation: Does the Model Understand the Rationale behind Logical Reasoning?

Context
Statistical records of crime rates probably often reflect as much about the motives and methods of those who compile or cite them [...]. The police may underreport crime [...] or overreport crime [...]. Politicians may magnify crime rates to [...]. Newspapers often sensationalize [...].

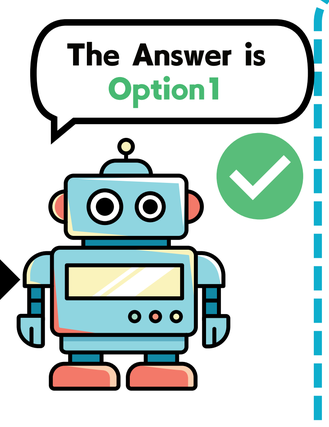
Question
The argument proceeds by

Option1
✓ citing examples in support of its conclusion.

⋮

Option4
✗ showing how evidence contradicts the conclusion.

Logical Multiple-Choice QA from ReClor (Yu+ 2020) (MainQ)



Our Scope

Why is Option1 Correct?
= Selective SubQ

SubQ

Why is Option4 NOT Correct?
= Eliminative SubQ

SubQ

Motivation

- Even if the model answers the question correctly, it is unclear whether the model also comprehends the rationale.
- The existing dataset (ReClor) doesn't evaluate the capability of understanding the rationale.

We construct a new dataset!

Our New Dataset: RULE

RULE

943 MainQ

Context
Question
Option1
⋮
Option4

3,003 SubQ

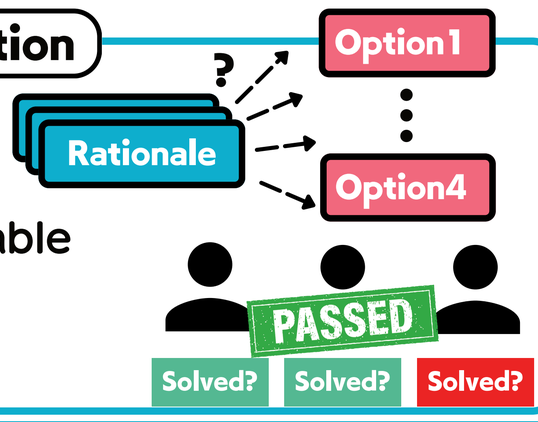
Context
Question for Option4
Rationale1 (✓ Because)
⋮
Rationale4 (✗ Because)

Dataset Design

- For the questions in ReClor (MainQ), we create multiple-choice questions asking for the reason why each option is correct/incorrect (SubQ)
- Selective SubQ asks why the correct op. should be selected. Eliminative SubQ asks why the incorrect op. should be eliminated.

Quality Validation

- Confirmed rationales are not too generic by matching test.
- Confirmed SubQs are answerable by ensuring human accuracy is ≧2/3.



How Well Current Models Answer SubQ Correctly?

Model	MainQ Acc.	SubQ Acc.	① Selective SubQ Acc.	② Eliminative SubQ Acc.
Five-Shot on ReClor				
FLAN-UL2	58.5$_{\pm 0.3}$	**65.5**$_{\pm 5.1}$	88.0$_{\pm 4.0}$	**57.6**$_{\pm 5.4}$
INSTRUCTGPT	71.8$_{\pm 1.0}$	65.3$_{\pm 1.8}$	88.4$_{\pm 2.5}$	57.1$_{\pm 1.5}$
INSTRUCTGPT + COT	67.8$_{\pm 0.5}$	63.2$_{\pm 2.1}$	88.5$_{\pm 2.5}$	54.2$_{\pm 2.8}$
LLAMA2 70B	**80.3**$_{\pm 0.4}$	60.0$_{\pm 2.6}$	**90.0**$_{\pm 1.1}$	49.4$_{\pm 2.9}$
HUMAN	**91.5**	**82.6**	**93.0**	**78.9**

① On Selective SubQs, the models' performances are close to that of human.

② On Eliminative SubQs, their performances significantly drop.

Can the Model Utilize Human-Crafted Rationales?

InstructGPT

	Input	Accuracy
	Context	72.2
①	+ Selective Rationale	91.4
	+ Eliminative Rationale	66.0
②	+ Both	89.6

① The model's performance boosts when given the selective rationale.

② Feeding the model the eliminative rationale degrades its performance.

Reference

- [Yu+ 2020] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A reading comprehension dataset requiring logical reasoning. *ICLR 2020.*