TellMeWhy: A Dataset for Answering Why-Questions in Narratives

Yash Kumar Lal[1], Nathanael Chambers[2], Raymond Mooney[3], Niranjan Balasubramanian[1]
Stony Brook University[1], US Naval Academy[2], University of Texas, Austin[3]
{ylal, niranjan}@cs.stonybrook.edu, nchamber@usna.edu, mooney@cs.texas.edu

Overview

- ➢ Humans can reason about why they do something
- ➢ Can a machine tell me why?
- ➢ Auto-extract action questions from ROCStories
- ➢ Solicit possible answers from 3 MTurk workers
- ➢ Collect judgments about answers’ validity
- ➢ **30k** questions, with **3** answers each
- ➢ Free-form generation task that models do bad on
- ➢ Advocate for human evaluation

Knowing why can help

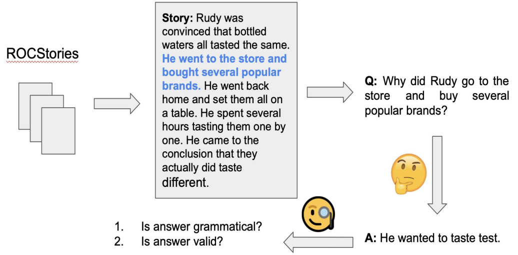
- ➢ Explain a character’s motivation.
- ➢ Visualize events in a narrative
- ➢ Understand plans and goals.

Story: Rudy was convinced that bottled waters all tasted the same. **He went to the store and bought several popular brands.** He went back home and set them all on a table. He spent several hours tasting them one by one. He came to the conclusion that they actually did taste different.

Q: Why did Rudy go to the store and buy several popular brands?

A: He wanted to taste test.

Dataset Creation



ROCStories

Story: Rudy was convinced that bottled waters all tasted the same. **He went to the store and bought several popular brands.** He went back home and set them all on a table. He spent several hours tasting them one by one. He came to the conclusion that they actually did taste different.

Q: Why did Rudy go to the store and buy several popular brands?

A: He wanted to taste test.

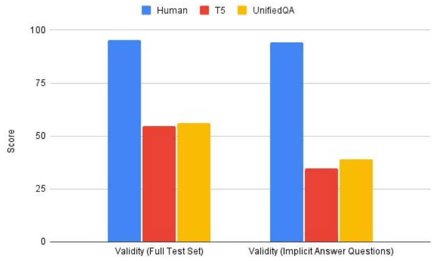
1. Is answer grammatical?
2. Is answer valid?

Characteristics

- ➢ Answers are not always in text - implicit answer (IA) questions
- ➢ IA questions prevent cheating - models can’t copy text to use as answers
- ➢ Diverse answers possible for each question
- ➢ Automatic evaluation is not enough

Split	# stories	# questions
Train	7558	23964
Dev	944	2992
Test	944	3099
Hidden Test	190	464
Total	9,636	30,519

Benchmarking



Conclusions

- ➢ Challenging why-question dataset
- ➢ Large LMs are not capable of answering why questions
- ➢ Harness for standardized human evaluation

Dataset Download

Scan this QR

Or, come find us at

bit.ly/sbu-tellmewhy