WikiWhy

Answering and Explaining Cause-and-Effect Questions

BACKGROUND

Previous QA benchmarks are susceptible to exploits such as context matching (extractive QA) or memorization. This hampers researchers' ability to measure and understand models' ability for reasoning. We introduce **WikiWhy**, a dataset containing 9,000+ "why" question-answer-rationale triples.

METHODS

Data Collection

1. Scrape passages from Wikipedia's curated "Good Articles" containing cause-effect pairs
2. MTurk stage 1†
   - **<input>** Passage **<output>** Cause & Effect + QA
3. MTurk stage 2†
   - **<input>** Cause & Effect **<output>** Explanation
4. Final Validation Stage: Verify and revise entries in the test/dev splits

† ran concurrently with worker-level validation of initial submissions

Baseline Experiments

Systems: Fine-tuned GPT-2, Few-Shot CoT GPT-3
Task Settings: (1) QA (2) Explanation (3) QA + Exp.

Automatic Evaluation

We introduce a new alignment based protocol for evaluating rationales against WikiWhy references.

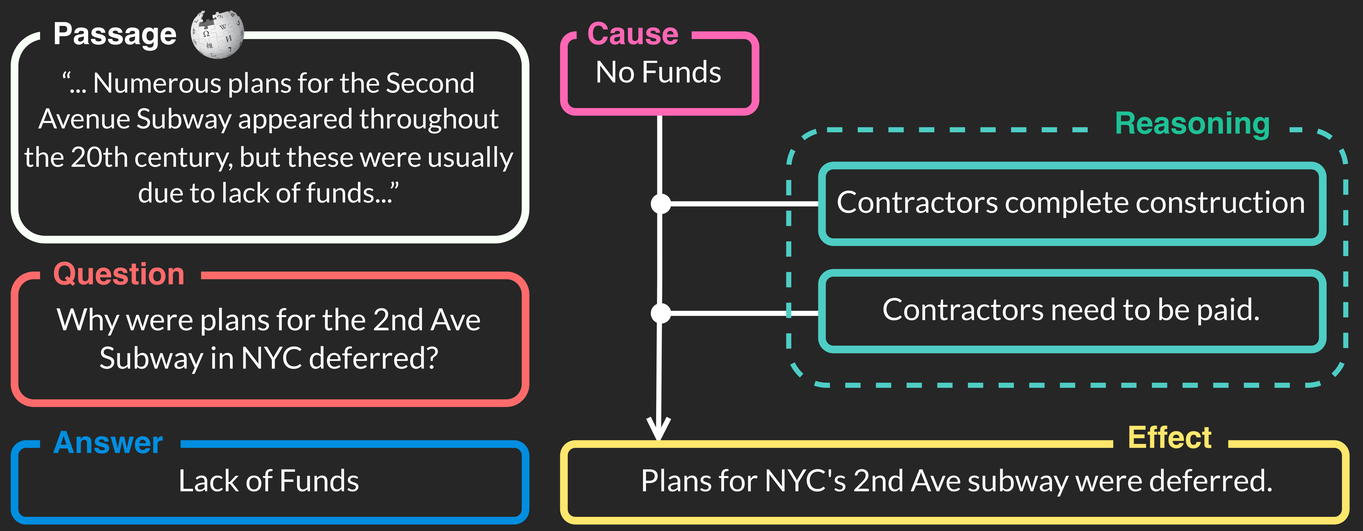
1. Unordered Alignment: compute similarity scores, threshold, then greedily align sentences from the predicted explanation against the reference.
2. Ordered Alignment: remove out-of-order matches using longest common subsequence.
3. Compute unordered and ordered F1 from each alignment respectively

Human Evaluation

We had a panel of students evaluate generated explanations compared to gold references.

Setting	Fine Grained Human Evaluation						
	Correctness	Concision	Fluency	Validity	Win (↑)	Tie	Lose (↓)
GPT-2: EO	0.100	0.880	0.860	0.520	0.040	0.040	0.920
GPT-3: EO	**0.660**	0.680	1.00	0.960	0.080	0.360	0.580
GPT-3: A&E	0.140	0.680	0.900	0.720	0.080	0.100	0.820

WikiWhy is a new benchmark for evaluating models' ability to ***explain*** **between cause & effect.**



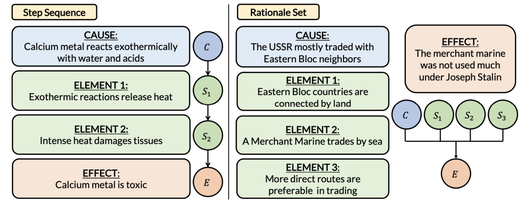
← read the paper

RELATED BENCHMARKS

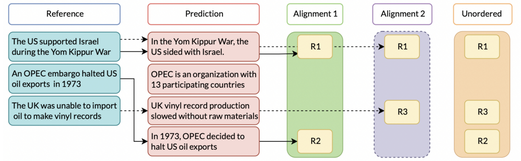
Dataset	Size	Answer Type	Explanation Type	Topics	Source
CoS-E[1]	9,500	MCQ	1-step	1	ConceptNet
eQASC[2]	9,980	MCQ	2-step	1	WorldTree
CausalQA[3]	24,000	Short	None	1	Yahoo Finance
EntailmentBank[4]	1,840	Short	Tree	1	WorldTree
WIKIWHY	**9,406**	**Short**	**Set/Chain**	**11**	**Wikipedia**

DATA COLLECTION PIPELINE

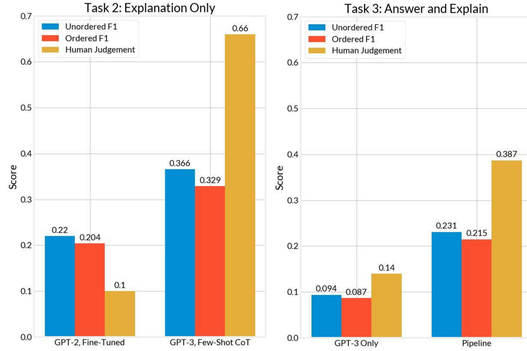
EXPLANATION TOPOLOGIES



AUTOMATIC EVALUATION



RESULTS



Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, William Wang

UC SANTA BARBARA