Possible Stories: Evaluating Situated Commonsense Reasoning under Multiple Possible Scenarios

Mana ASHIDA* (Yahoo Japan Corporation), Saku SUGAWARA (National Institute of Informatics)

Contact: maashida@yahoo-corp.jp, saku@nii.ac.jp *Work done while at Tokyo Metropolitan University.

Goals

- **Proposing a task and a benchmark dataset to evaluate machines' commonsense reasoning ability under many conditions.**
- **Testing how well the current state-of-the-art models perform compared to humans.**
- **Creating a challenging multiple choice QA benchmark where models cannot use statistical patterns or simple heuristics to answer correctly.**

Background

Evaluating Machines' Commonsense Reasoning

Many benchmarks have been proposed to evaluate machines' commonsense reasoning ability. Most of them challenge models to answer the most plausible option under ***any*** circumstances (therefore, **not situated** in a particular context) among multiple options to questions about facts, causality, logic, and so on. However, there happen multiple plausible and possible consequences to a single context in reality.

Plausible Options Changing across Situations

Natural language understanding benchmarks incorporating several possibilities to one question or one context, along with when it becomes plausible are a handful.

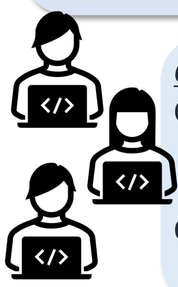
- SituatedQA (Zhang and Choi, 2021): objective facts across time and place.
- Moral Stories (Emelin et al., 2021): human behavior either normative or divergent.

Our Dataset: Possible Stories

Original Context
Cindy was planning to grow a lot of vegetables this year. She planted vegetable seedlings in her garden. Cindy knew there were hungry groundhogs in the area. She put up a short fence around her garden to protect it.

Original Ending
A: The groundhogs climbed over the fence and ate her seedlings.

Alternative Ending Collection
B: All of the ground animals were kept out, but something was still eating her vegetables.
C: She put spikes on the fence to avoid groundhogs and it worked.
D: No groundhogs climbed over the fence and Cindy had a good harvest in the fall.



Question Writing
Q1: Which one of the following is most likely to happen after this if there were other hungry animals? → **Option B**
Q2: What would be the most positive outcome for Cindy's crops? → **Option D**

Possible Stories contains 4.5k questions to 1.3k passages (avg. 3.45 questions/passage), entirely written by crowdworkers on Amazon MTurk.

Context: ROCStories (Mostafazadeh et al., 2016)

ROCStories is a collection of five-sentence stories on everyday situations.

Endings

Workers create possible alternative endings to the fifth sentence of the context (the original ending).

Questions

Workers generate questions with only one option as the answer; namely, one of the endings becomes the most plausible among the others.

Quality Control

Running validation sessions to ensure 1) questions have only one clear answer and 2) the dataset does not contain offensive words & unfair descriptions of particular groups of people.

Example

Context: Jan checked to make sure no one was around. Her two older brothers had been sneaking around the garden lately. Being a curious child, Jan wanted to know what they were up to. She carefully opened the door to her brother's room.

Questions:

1. If Jan smelled pleasant aromas and felt fresh air in the room, what did she likely discover? (condition) ➡️ Answer: **Option C**
2. What was the likely outcome if Jan was left still feeling clueless about what her brothers had been up to? (character) ➡️ Answer: **Option D**
3. Which outcome is the most unlikely to occur in reality? (fiction) ➡️ Answer: **Option B**
4. Which would be particularly unpleasant for Jan if she suffers from acute arachnophobia? (character) ➡️ Answer: **Option A**

Options:

A. Inside the back of their closet, she found several jars with spiders.
B. There was a strange looking alien peeking out of a corner with fearful eyes.
C. They had taken plants from the garden and moved them to their room.
D. The door slammed shut on her face as the cameras alerted the brothers of an intruder.

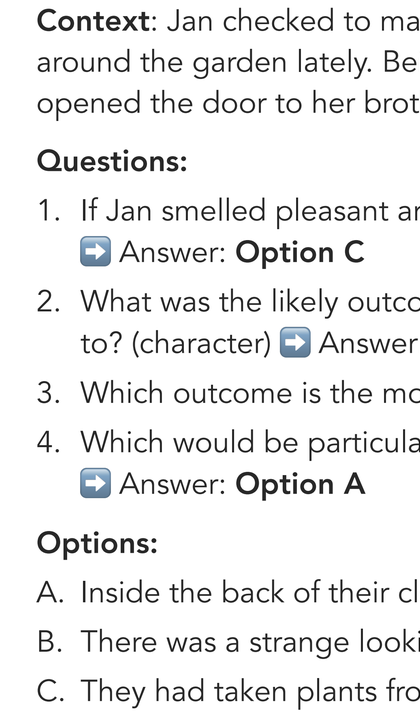
Experiments and Result

Model	Unsupervised	w/o passage	w/o question
DeBERTa-large v3	**60.2**	58.1	21.8
RoBERTa-large	50.5	50.3	21.5
Perplexity	30.4	35.4	26.4
Semantic similarity	37.3	47.1	28.8
Entailment	23.1	—	—
Human	**92.5**	—	—

Model and human performances on our dataset

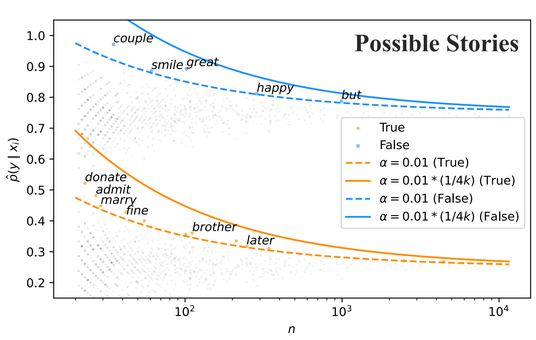
DeBERTa (He et al., 2021) and RoBERTa (Liu et al., 2019) models are fine-tuned on RACE (Lai et al., 2017), which is a large-scale multiple-choice reading comprehension dataset of middle and high school English exams and has passages and questions on various topics.

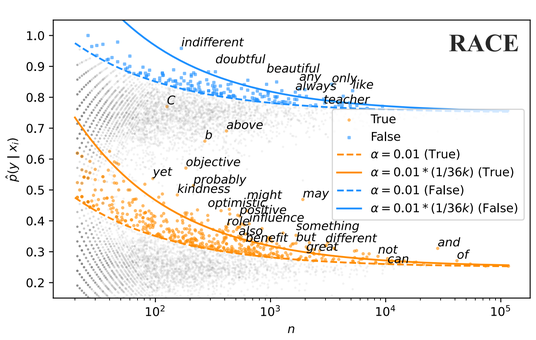
Analysis of Reasoning Types and Annotation Artifacts



Reasoning types

Proposing nine categories of reasoning types and annotated Possible Stories, CosmosQA (Huang et al., 2019), and RACE. Some of the reasoning types are only present in Possible Stories.





Annotation Artifacts (Gururangan et al., 2018; Gardner et al., 2021)

Annotation artifacts are statistical patterns between inputs and output labels found in crowdsourced datasets. Significantly fewer tokens are found compared with RACE or CosmosQA upon statistical test.

Conclusion

- We propose a situated commonsense reasoning task and create a multiple-choice QA dataset. (accessible at https: //github.com/nii-cl/possible-stories.)
- We discover that current strong pretrained language models struggle to solve our task when training data are unavailable.
- We show that our dataset contains minimal annotation artifacts in the answer options and has many challenging questions that require counterfactual reasoning and understanding characters' motivations and reactions, readers' perceptions, and fictional information.

COLING 2022