COLING 2022
Oct. 12-17, 2022 @ Gyeongju, South Korea

Possible Stories: Evaluating Situated Commonsense Reasoning under Multiple Possible Scenarios

Mana Ashida (Yahoo Japan Corporation)* *maashida@yahoo-corp.jp*
Saku Sugawara (National Institute of Informatics) *saku@nii.ac.jp*

*work done while at Tokyo Metropolitan University.

Background

Commonsense reasoning is a fundamental intelligence acquired with humans, and researchers are interested in if it is learned by the models.

Many benchmarks have been proposed to measure machine’s commonsense reasoning ability, but the current state-of-the-art models’ performance is comparable to human.

Background

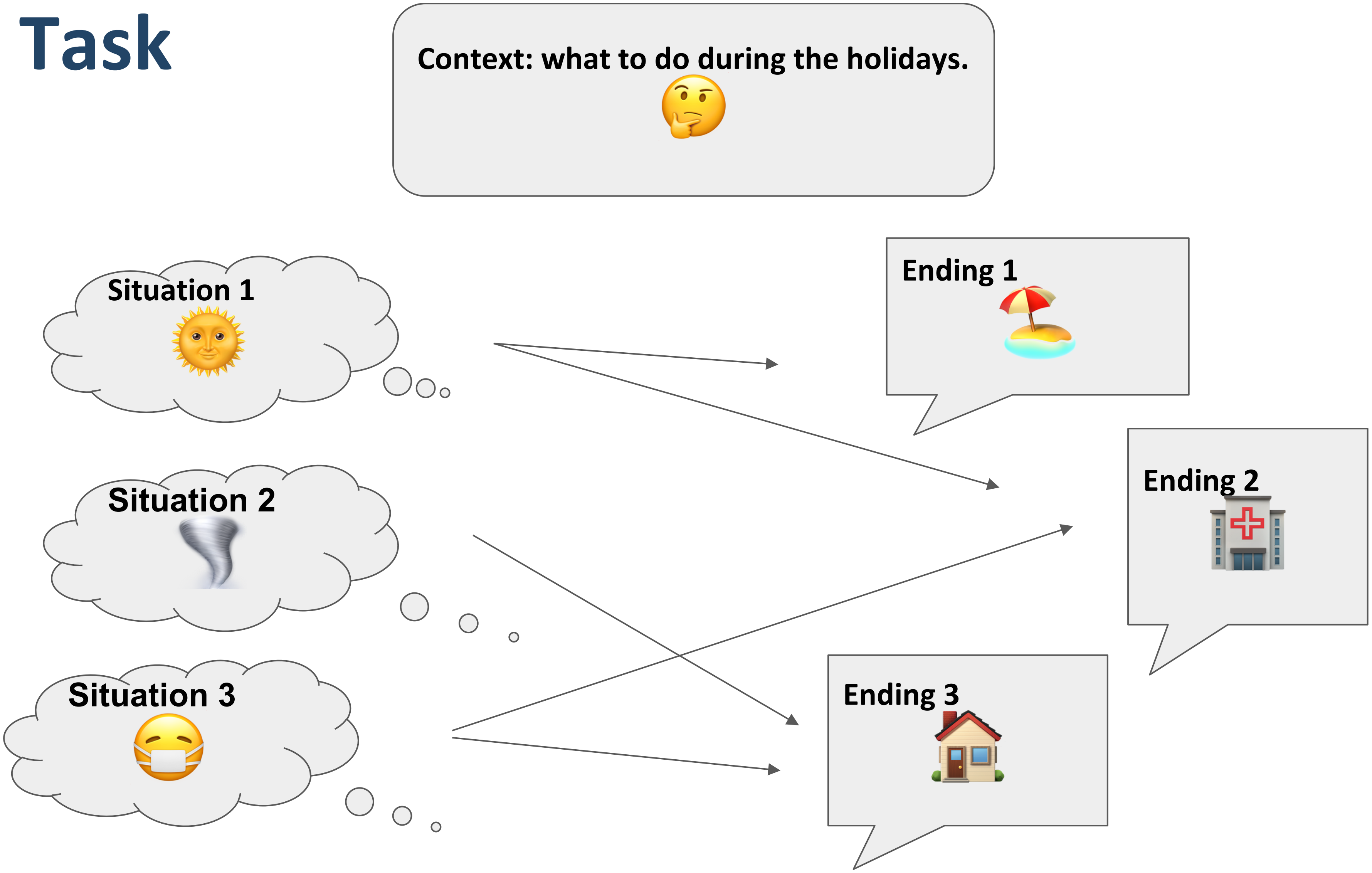
Natural language understanding (NLU) benchmarks incorporating several possibilities to one question, along with when it becomes plausible are still rare.

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

Our Goals

- Creating a resource to evaluate commonsense reasoning **under many different types of conditions**.
- Probing how well the current SoTA models perform, compared with human.

Situated Commonsense Reasoning Task



Context: what to do during the holidays.

Situation 1

Situation 2

Situation 3

Ending 1

Ending 2

Ending 3

Dataset Creation

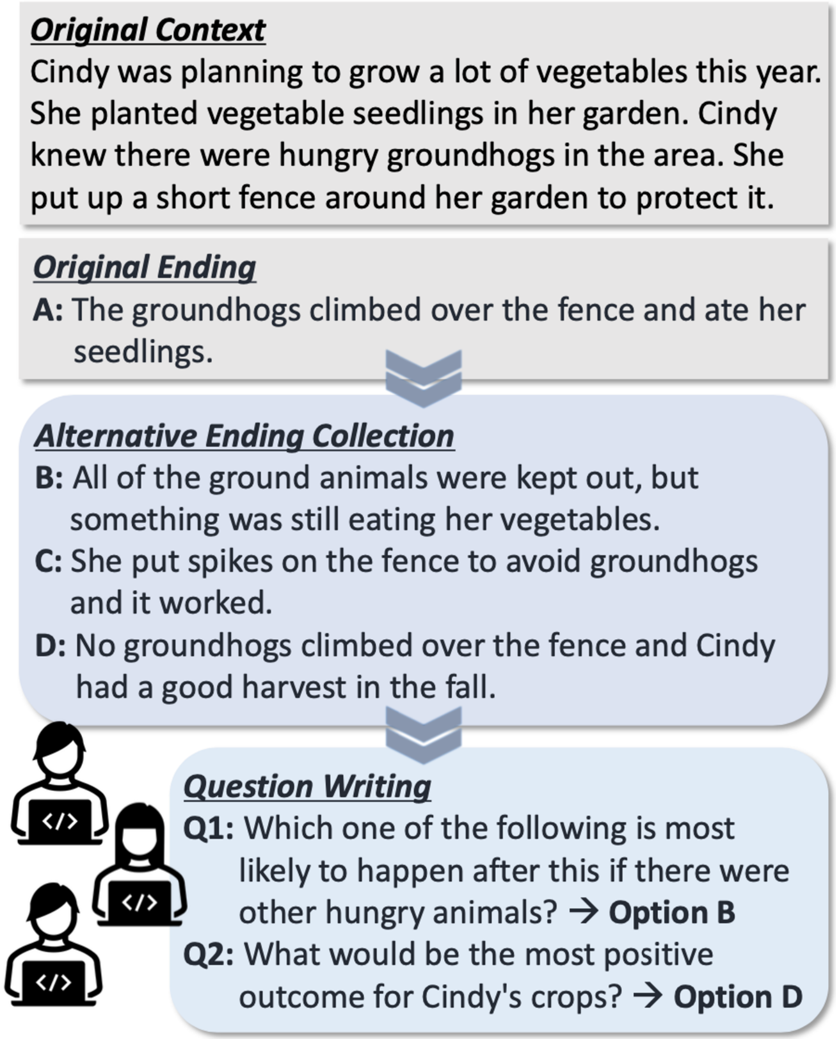
ROCStories (Mostafazadeh et al., 2016)

A collection of short stories consists of five sentences.

- Beginning and ending are clear.
- Not too generic, not too specific.
- Stories on everyday situations.

Dataset Creation

1. Alternative Ending Collection
2. Question Writing
3. Validation
   a. Question-Answer
   b. Content



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**

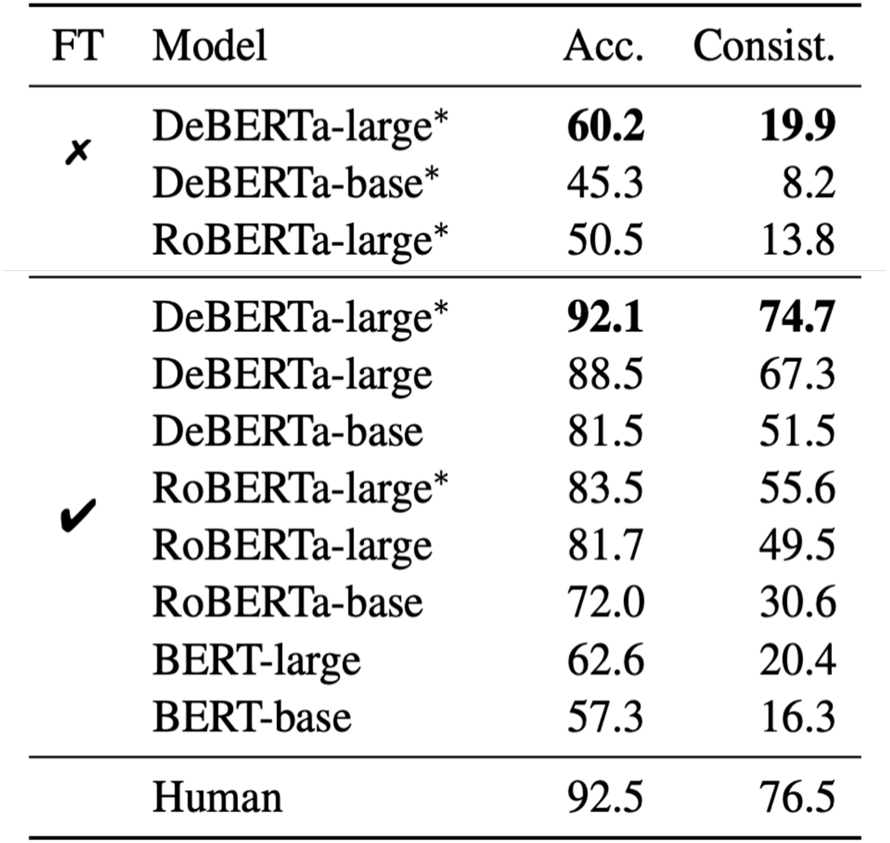
Possible Stories Overview

- 4.5k questions to 1.3k passages, 3.45 questions per passage.
- The average # of tokens:
  - 14.2 for question, 15.3 for options

    → longer than RACE (Lai et al., 2017) or CosmosQA (Huang et al., 2019)
- Dataset split:
  - train (75%), dev (10%), test (15%)
  - Dev and test set contain the examples generated by the workers who perform very well.

Experiments

- *unsupervised*
  - fine-tune with RACE
- *supervised*
  - fine-tune with Ours



FT	Model	Acc.	Consist.
✗	DeBERTa-large*	**60.2**	**19.9**
	DeBERTa-base*	45.3	8.2
	RoBERTa-large*	50.5	13.8
✔	DeBERTa-large*	**92.1**	**74.7**
	DeBERTa-large	88.5	67.3
	DeBERTa-base	81.5	51.5
	RoBERTa-large*	83.5	55.6
	RoBERTa-large	81.7	49.5
	RoBERTa-base	72.0	30.6
	BERT-large	62.6	20.4
	BERT-base	57.3	16.3
	Human	92.5	76.5

Model and human performances on our dataset. (∗) indicates that the model is fine-tuned on RACE.

Case Study

Context: A family going on a trip in the summer and made new friends.

Options:

A. They kept in touch with their friend even after they went home.
B. At the end of the day the kids got into a fight with each other and were happy to leave. ✅
C. The Smith’s decided they’d visit a new beach every year, and they made tons of new friends.
D. They went home though and the kids never saw their friend again.
⏪ DeBERTa-large

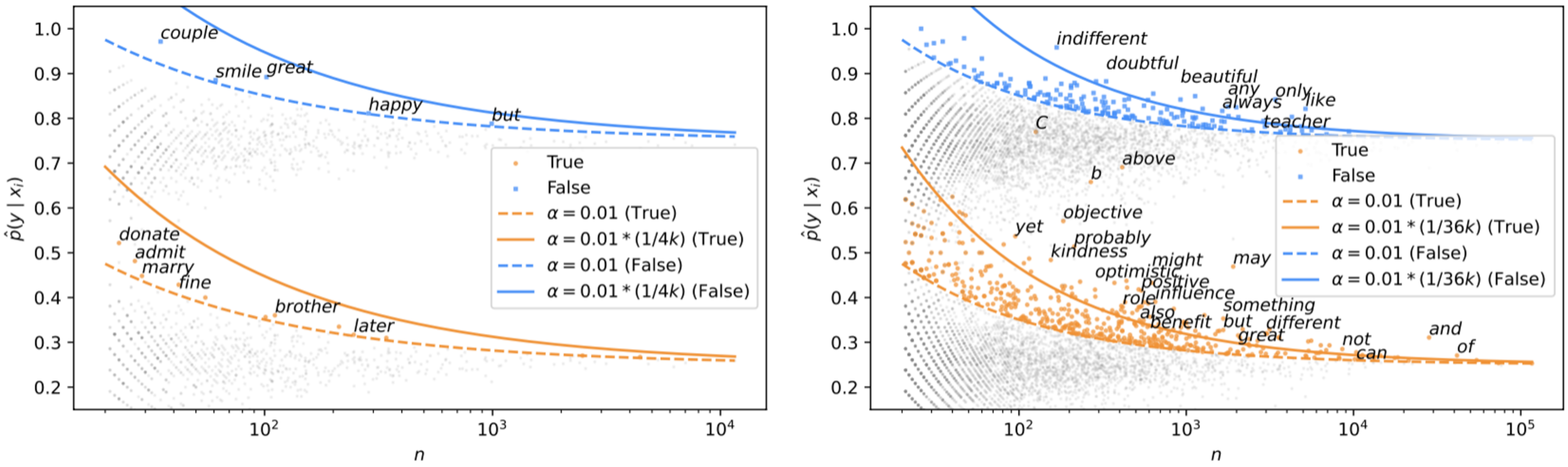
***Question**: <u>Which ending involves the most conflict?</u>*

Analysis of Annotation Artifacts

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

Statistical patterns between inputs and output labels found in the

(

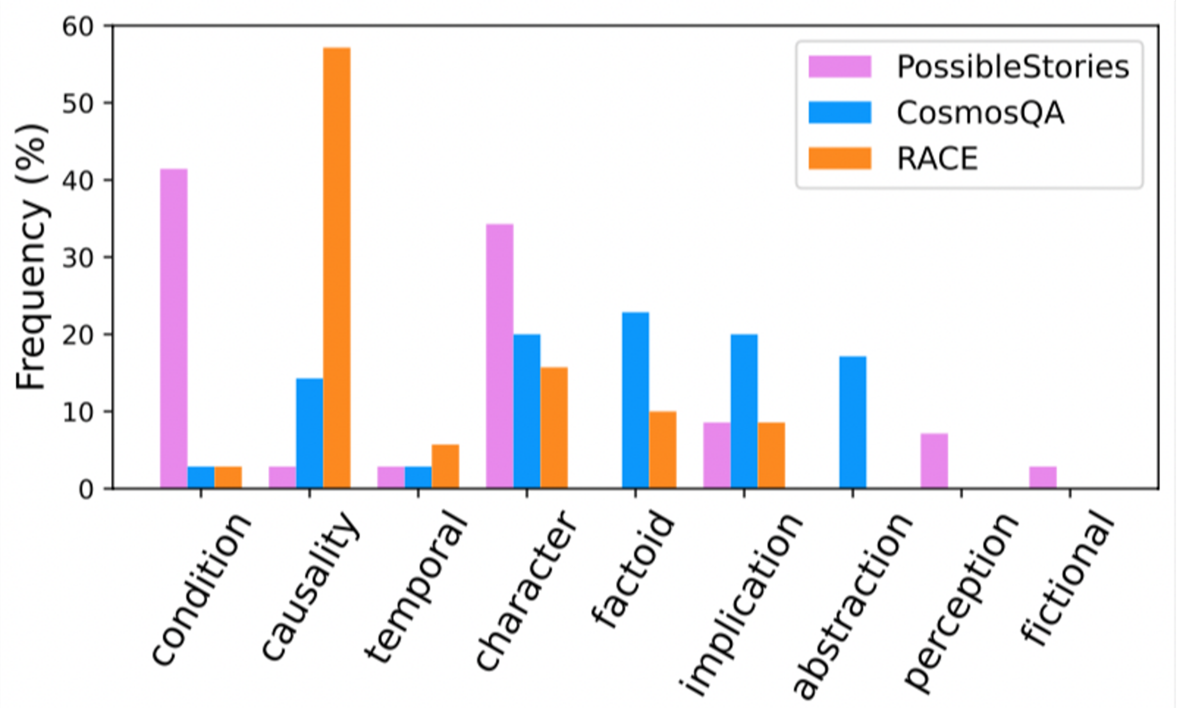


(a) Possible Stories (ours)

(b) RACE

Analysis of Reasoning Types

- Classifying questions into nine reasoning types.
- Possible Stories dataset contains questions with following types which are often absent in existing datasets.



Conclusions

- 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 an understanding of characters' motivations and reactions, readers' perceptions, and fictional information.