EECS 595 Project Presentation

An Attempt to Improve the Baseline Performance on TRIP

Zhixuan Chen
Yichen Zhong
Team 11

Content

- Introduction & Motivation (2min)
- Relation to Previous Work (0.5min)
- Methods (2min, 1 for ART, 1 for DistilBert)
- Results (1min)
- Future Plan (0.5min)

Introduction - Tiered Reasoning for Intuitive Physics

Why TRIP?

- LLMs perform well on lots of NLP tasks, but does the great performance come from proper reasoning process?
- TRIP is proposed to figure out the LLMs’ intermediate reasoning process.

Introduction - TRIP Dataset

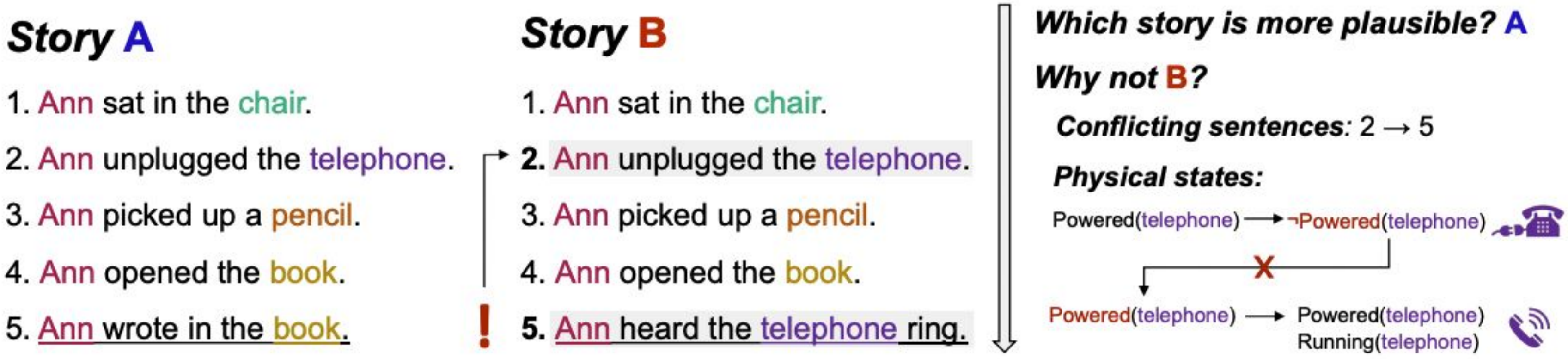


Figure 1: Story pair from TRIP, along with the tiers of annotation available to represent the reasoning process.

Introduction - Reasoning System for TRIP

Physical State Classification → Conflicting Sentence Detection → Story Choice classification

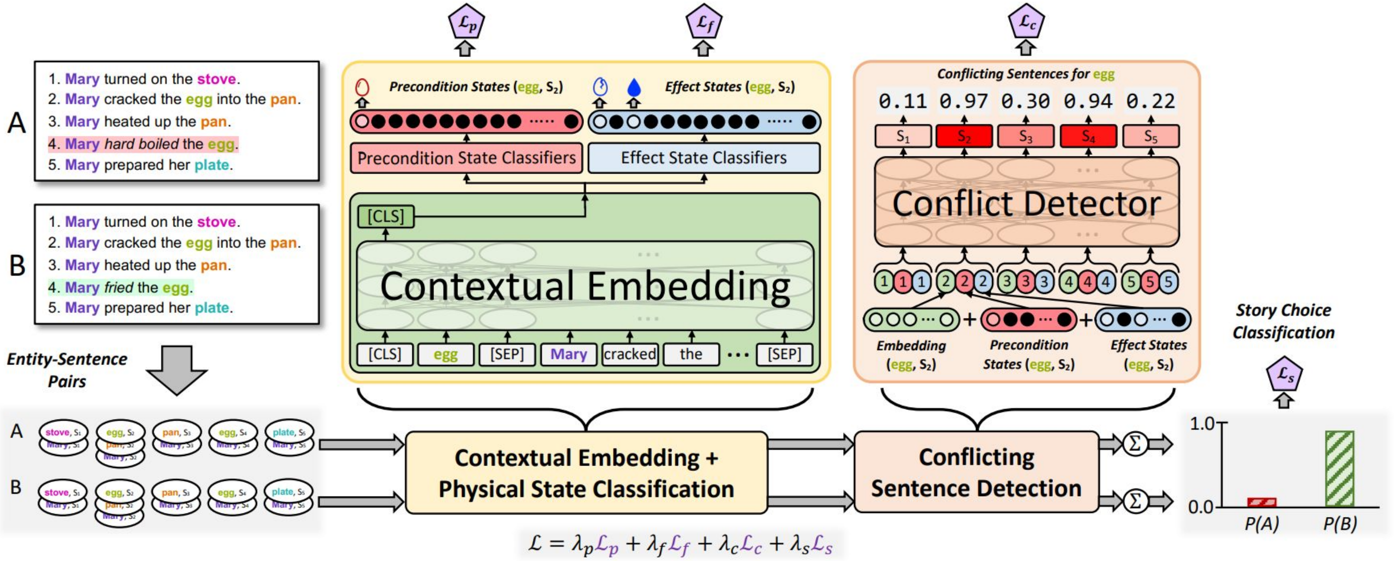


Figure 2: Proposed tiered reasoning system with loss functions $\mathcal{L}_p$ for precondition state classification, $\mathcal{L}_f$ for effect state classification, $\mathcal{L}_c$ for conflicting sentence detection, and $\mathcal{L}_s$ for story choice classification. The model is trained end-to-end by optimizing the joint loss $\mathcal{L}$, a weighted sum of these loss functions.

Introduction - TRIP Results

Evaluation Metrics

- **Accuracy**

  The plausible story is correctly identified

- **Consistency**

  The plausible story is correctly identified + The conflicting sentence pair for the implausible story is correctly identified (28% at best)

- **Verifiability**

  The plausible story is correctly identified + The conflicting sentence pair for the implausible story is correctly identified + Underlying physical states that contribute to the conflict are correctly identified (10.6% at best)

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
random	47.8	11.3	0.0
All Losses			
BERT	**78.3**	2.8	0.0
RoBERTa	75.2	6.8	0.9
DeBERTa	74.8	2.2	0.0
Omit Story Choice Loss $\mathcal{L}_s$			
BERT	73.9	**28.0**	9.0
RoBERTa	73.6	22.4	**10.6**
DeBERTa	75.8	24.8	7.5
Omit Conflict Detection Loss $\mathcal{L}_c$			
BERT	50.9	0.0	0.0
RoBERTa	49.7	0.0	0.0
DeBERTa	52.2	0.0	0.0
Omit State Classification Losses $\mathcal{L}_p$ and $\mathcal{L}_f$			
BERT	75.2	17.4	0.0
RoBERTa	71.4	2.5	0.0
DeBERTa	72.4	9.6	0.0

Methods

- Previous work: TRIP Dataset & related Paper
- Our goal: Improve Baseline Performance of the TRIP Dataset
- Our work:
    - Reproduce the result of the TRIP paper
    - Apply Transfer Learning from a new dataset ART to train TRIP dataset (new!)
    - Integrate other evaluation methods (e.g. DistilBert) to train TRIP dataset (new!)
    - Compare results and discussion

Approach I: Transfer Learning using ART

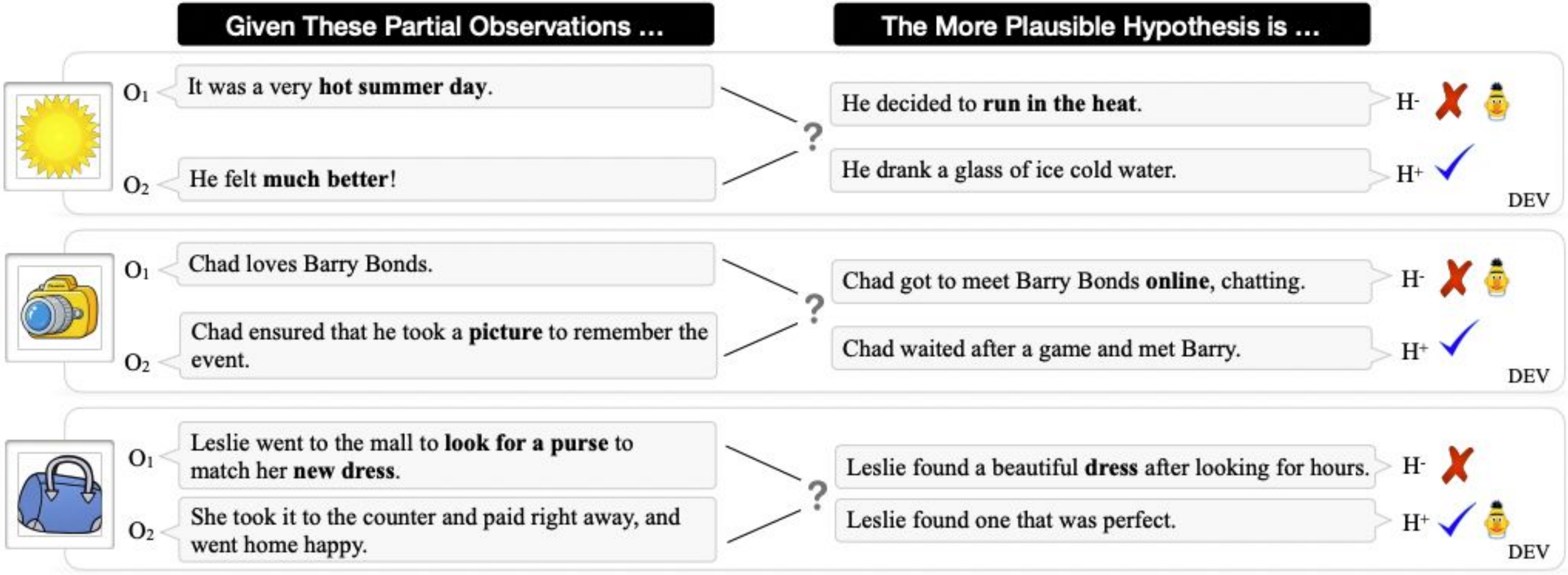
Reason:

- ART is similar to the TRIP dataset
- ART size is much larger than TRIP
- May generate better results

Approach:

- Load and manipulate the ART dataset
- Train, Validate and Test the dataset
- Hyper-parameter Tuning to get the best performance parameters

Approach I: Transfer Learning using ART



Approach II: Evaluation Performance using DistilBert

Reason:

- Original TRIP paper only evaluate using BERT, RoBERTa, and DeBERTa
- Other pretrain model might work better
- DistilBert is a relatively light model, convenient for testing

Result I: Fine-tune on ART

Bert-base-uncased with ART

Batch size = 20; learning rate = 1e-5	52.02%
Batch size = 20; learning rate = 5e-5	53.13%
Batch size = 15; learning rate = 1e-5	54.44%
Batch size = 15; learning rate = 5e-5	**54.57%**
Batch size = 10; learning rate = 1e-5	53.00%
Batch size = 10; learning rate = 5e-5	52.74%

Result II: Fine-tune on TRIP

Model Type	Loss Type	Accuracy	Consistency	Verifiability
BERT-base-uncased with TRIP (ART finetuned)	All losses	73.79%	1.14%	0.28%
	Omit story choice loss	72.93%	19.09%	3.70%
	Omit conflict detection loss	41.60%	0.00%	0.00%
	Omit state classification losses	75.78%	3.13%	0.00%
BERT-base-uncased with TRIP (without ART)	All losses	78.92%	0.85%	0.28%
	Omit story choice loss	71.79%	18.80%	4.27%
	Omit conflict detection loss	42.45%	0.00%	0.00%
	Omit state classification losses	76.64%	8.26%	0.00%
DistilBERT with TRIP (without ART)	All losses	78.92%	1.14%	1.14%
	Omit story choice loss	72.57%	17.38%	8.63%
	Omit conflict detection loss	38.46%	0.00%	0.00%
	Omit state classification losses	70.37%	3.13%	0.00%

Future Work

- We have got results, but have not done analysis yet
- For Future Work:
  - Compare the result we generated and the paper benchmarks
  - Conduct performance analysis about the result
  - Discuss the differences, improvements, and potential reasons among the results