# CIS²: A Simplified Commonsense Inference Evaluation for Story Prose

Bryan Li, Lara J. Martin, and Chris Callison-Burch

Contact: bryanli@seas.upenn.edu

Scan to read full paper

## INTRODUCTION

- *Contextual Commonsense Inference (CCI)*: inferring causal relations between events in text
- GLUCOSE [1] dataset is valuable, but GLUCOSE task conflates CCI and generation (NLG)
- Finetuning T5 [2] LM → model relies on NLG
- Evaluating with BLEU → partial matches count

| Parameter | Text |
|---|---|
| Story | Fred woke up late. He just missed his bus. He then went to his mom's room. His mom then drives him to school. He makes it to first class on time. |
| Selected Sentence (X) | Fred woke up late. |
| Dimension | 6 |
| Specific Rule | Fred wakes up late >Causes/Enables> Fred misses his bus |
| General Rule | Someone$_A$ wakes up late >Causes/Enables> Someone$_A$ misses Something$_A$ |

Input

Output

**Transformers trained on *commonsense inference* tasks will rely on their generation ability, if given the chance, which inflates performance.**

## METHODS

### 1. Diagnose extent to which CCI and NLG are conflated in the GLUCOSE task

| Task | Input |
|---|---|
| ORIGINAL | 1: My mother told me to fix the car. I was unable to do this right away. *** I could not find my tools. *** I looked everywhere for them. It turns out they were stolen the night before. |
| HISTORY | 1: My mother told me to fix the car. I was unable to do this right away. |
| MASK X | My mother told me to fix the car. I was unable to do this right away. <masked> I looked everywhere for them. It turns out they were stolen the night before. |
| HISTORY+X | 1: My mother told me to fix the car. I was unable to do this right away. *** I could not find my tools. *** |

**Output:**
They were stolen the night before >Causes/Enables> **I could not find my tools **** Something$_A$ is stolen >Causes/Enables> Someone$_A$ cannot find Something$_A$

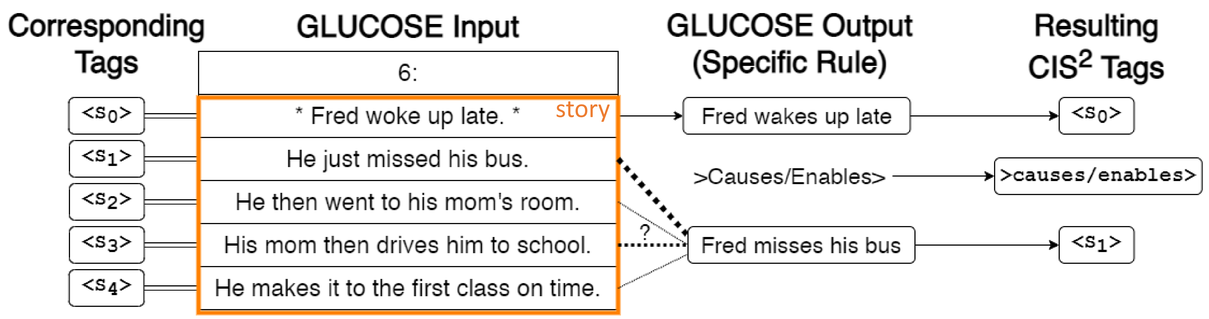
### 2. Suggest an alternative evaluation: CIS²

**(Contextual Commonsense Inference in Sentence Selection)**

Abstract away from NLG, only consider causal links between sentences

1. *Convert* GLUCOSE to CIS² labels using SBERT[3] similarity metric
2. *Train* T5 on CIS² converted data to compare to GLUCOSE models
3. *Evaluate* using **exact match** CIS² accuracy

**CONVERSION**



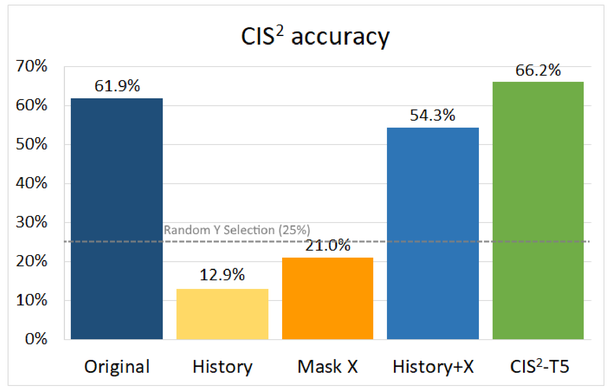

**Output:** <s$_4$> >Causes/Enables> <s$_2$>

## RESULTS

### 1. Diagnose where CCI is being conflated with generation

| Task | Specific Rule (SacreBLEU) | General Rule (SacreBLEU) |
|---|---|---|
| ORIGINAL | 70.7 | 66.2 |
| HISTORY | 35.9 | 50.4 |
| MASK X | 41.6 | 49.6 |
| HISTORY+X | 68.3 | 65.5 |

### 2. Suggest alternative evaluation: CIS²




## DISCUSSION

Input

#6: *Fred woke up late.* He just missed his bus. He then went to his mom's room. His mom then drives him to school. He makes it to first class on time.

position of selected sentence

Output

Fred wakes up late >Causes/Enables> Fred misses his bus

** Someone$_A$ wakes up late >Causes/Enables> Someone$_A$ misses Something$_A$

- T5 will rely on generation if it is an option.
- CCI tasks need to put less emphasis on generation in order to evoke the "reasoning" capabilities of Transformers.

[1] Nasrin Mostafazadeh et al. GLUCOSE: GeneraLized and COntextualized story explanations. In *EMNLP*, 2020.
[2] Colin Raffel, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.
[3] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP*, 2019.