Get Your Vitamin C!

Robust Fact Verification with Contrastive Evidence

Tal Schuster, Adam Fisch, and Regina Barzilay

MIT CSAIL

Overview

- **VitaminC dataset:** contains almost 500K claim-evidence pairs with pairs contrastive evidence, based on Wikipedia factual revisions.
- Increases sensitivity of fact verification models to changes in the evidence.

Example: Our model correctly predicts the following claim as False. Most models mistakenly predict it as True, even when provided with refuting evidence from Wiki:

*The 2020 Olympics were held in **2020***

Evidence

2020 Summer Olympics

From Wikipedia, the free encyclopedia

[…] scheduled to be held from 23 July to 8 August **2021**.

TOKYO 2020

- There are over 6k revisions to Wikipedia every hour. We collect ~100K revisions and add synthetic ones to cover a wide range of topics and categories, including COVID-19 related claims.
- In addition to improving the robustness of classifiers, we formulate and evaluate new tasks in the ecosystem of fact verification.

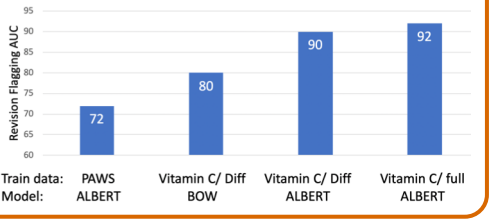
Factual Revision Flagging

- Many revisions are not factual. We are interested in identifying the ones that modify an underlying fact. **Example:**

The outbreak was first identified in December 2019 and recognized [...]. s_{t-1}

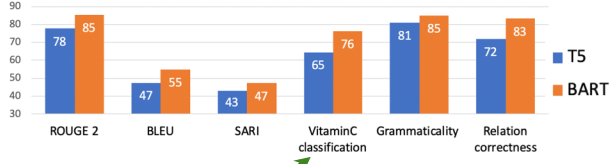
The outbreak was first identified in 17 November 2019 and recognized [...]. s_t

Results:



Factually Consistent Generation

- **Automatic revision:** Update Wikipedia
- **Claim extraction:** Distil the factual change



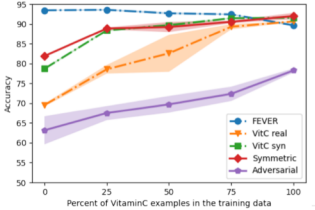
- Our fact verification classifier can estimate the consistency of the output and catch hallucinations common in LMs like GPT-3.

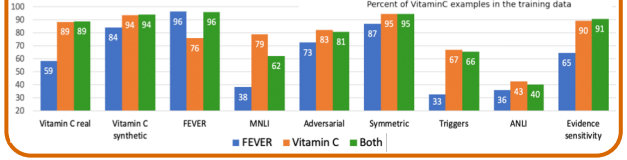
Robust Fact Verification

- A model should predict the relation of each version of Wiki right, so it will be sensitive to future updates of information.

Claim: *COVID-19 was identified before Dec*

- VitaminC-trained models are more robust
- Helps with both fact verification and natural language inference (MNLI).





Word-level Rationales

- Word-level rationales in the evidence can increase the trust of users in the model's predictions.