AVeriTeC: A Dataset For Real-World Claim Verification with Evidence from the Web

UNIVERSITY OF CAMBRIDGE

Michael Schlichtkrull*
Zhijiang Guo*
Andreas Vlachos

Overview

- We introduce a new dataset of 4,568 claims for **A**utomated **Veri**fication of **Te**xtual **C**laims (AVeriTeC).
- We employ crowdworkers to turn fact-checking articles from journalists into sequences of open-domain QA problems.

Split	Train	Dev	Test
Claims	3068	500	1000
Questions / Claim	2.60	2.57	2.57
Reannotated (%)	28.1	24.4	25.1
End date	25-08-2020	31-10-2020	22-12-2021
Labels (S / R / C / N) (%)	27.6 / 56.8 / 6.4 / 9.2	24.4 / 61.0 / 7.6 / 7.0	25.5 / 62.0 / 6.3 / 6.2

Previous Datasets

Group 1: *Synthetic, purpose-made claims*, high-quality evidence.

Group 2: Real claims, but evidence (if at all given) *is* the AFC article.

Group 3: Real claims, evidence from the web, but *insufficient*.

Group 4: Real claims, evidence from the web, but *temporally leaked*.

For more discussion, see also:

- Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation. Glockner et al., EMNLP 2022.
- Varifocal Question Generation for Fact-checking. Ousidhoum et al., EMNLP 2022.

Claim: *The USA has succeeded in reducing greenhouse emissions in previous years.*
Date: **2020.11.2**

Q1: What were the total gross U.S. greenhouse gas emissions in 2007?
A1: In 2007, total gross U.S. greenhouse gas emissions were 7,371 MMT.

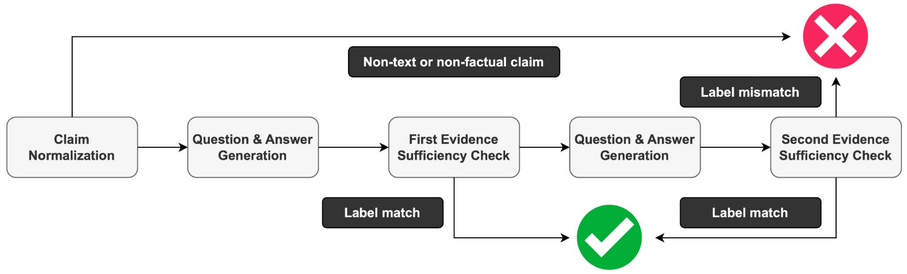
Q2: When did greenhouse gas emissions drop in the USA?
A2: In 2017, total gross U.S. greenhouse gas emissions were 6,472.3 MMT, or million metric tons, carbon dioxide.

Q3: Did the total gross U.S. greenhouse gas emissions rise after 2017?
A3: Yes. After 3 years of decline, US CO2 emissions rose sharply last year. Based on preliminary power generation, natural gas, and oil consumption data, we estimate emissions increased by 3.4% in 2018.

Verdict: **Conflicting Evidence/Cherrypicking.**
Justification: *It is true they did reduce emissions however they have now increased again. It is unknown exactly what years are being referred to.*

AVeriTeC

- Guarantees *checkworthiness, sufficiency,* and *temporal unleakedness.*
- Represents retrieval and reasoning as *question-answer pairs*, a natural format that allows reuse of models from other tasks.
- Includes *justifications* that explain how question-answer pairs lead to verdicts.
- Is available at https://github.com/MichSchli/AVeriTeC.
- Will be the shared task at FEVER @ EMNLP 2024.



Baseline

Model	Q only	Q + A	Veracity @ .25	Justifications @ .25
No search	0.19	0.11	0.02	0.01
Gold evidence	1.00	1.00	0.49	0.28
AVeriTeC -BLOOM-7b	0.26	0.21	0.15	0.07
gpt-3.5-turbo	0.29	0.16	0.10	0.04

- Our model: BLOOM for QG, Google + BM25 + BLOOM + BERT for QA, BERT for verdicts, BART for justifications.
- No search: same model, but QA component always outputs *"no answer could be found"*.
- Gold evidence: same model, but generated QA pairs are replaced with gold QA pairs.
- Our baseline performs reasonably, but there is room for improvement (maybe your model?)
- ChatGPT is often right about the verdict, but hallucinates fake evidence – this is not enough for real-world fact-checking!

European Commission
erc