Ask to Know More: Generating Counterfactual Explanations for Fake Claims

Shih-Chieh Dai* (The University of Texas at Austin), **Yi-Li Hsu*** (National Tsing Hua University, Academia Sinica), **Aiping Xiong** (The Pennsylvania State University), **Lun-Wei Ku** (Academia Sinica)

*Equally contribution sjdai@utexas.edu

Paper

Code

Introduction

We aimed to generated **counterfactual explanations** for why a piece of fake news is fake.

Research Questions:

1. How can we generate a good counterfactual explanation for a given fake claim?
2. Do different types of counterfactual explanations (i.e., **affirmative**, **negative**, and **mixed**) vary in best explaining why a piece of news is fake?
3. How do counterfactual explanations best explain why a piece of news is fake compared to other state-of-the-art explanations?
4. Does an individual's **familiarity** (familiar vs. unfamiliar) with misinformation impact the effectiveness of counterfactual? explanations?

Counterfactual Explanation

Definition: The result of doing something that is counter to fact. [1]

False Claim: **Istanbul's population has increased by 400 percent since the 1950s.**

Evidence: Istanbul's population has increased tenfold since the 1950s , as migrants from across Anatolia have moved in and city limits have expanded to accommodate migrants from across Anatolia .

Questions and Answers:

Q: What has increased by 400 percent since the 1950s?

A: Istanbul's population has increased tenfold since the 1950s.

Q: What is the largest city in Turkey?

A: Istanbul is the largest city in turkey.

Q: How much has Istanbul's population increased since the 1950s?

A: Tenfold

Best Answer: Tenfold

Declarative Sentence:
Istanbul's population has increased tenfold since the 1950s.

Counterfactual Explanation:
If we say 'Istanbul's population has increased tenfold since the 1950s' instead of 'Istanbul's population has increased by 400 percent since the 1950s', the claim would be correct.

Counterfactual formats [2]**:**

Affirmative (CF-A): "If we were to say Si instead of Fi , the claim would be correct."

Negative (CF-N): "If we were to say not Ci but instead Si , the claim would be correct."

Mixed (CF-M): "If we were to say NCi and/but say Fi, the claim would be correct."

Ci: Claim, Si: Declarative Sentence

Fi: smallest change needed to Ci to flip the reader's opinion.

NCi: the negation of the false claim

References

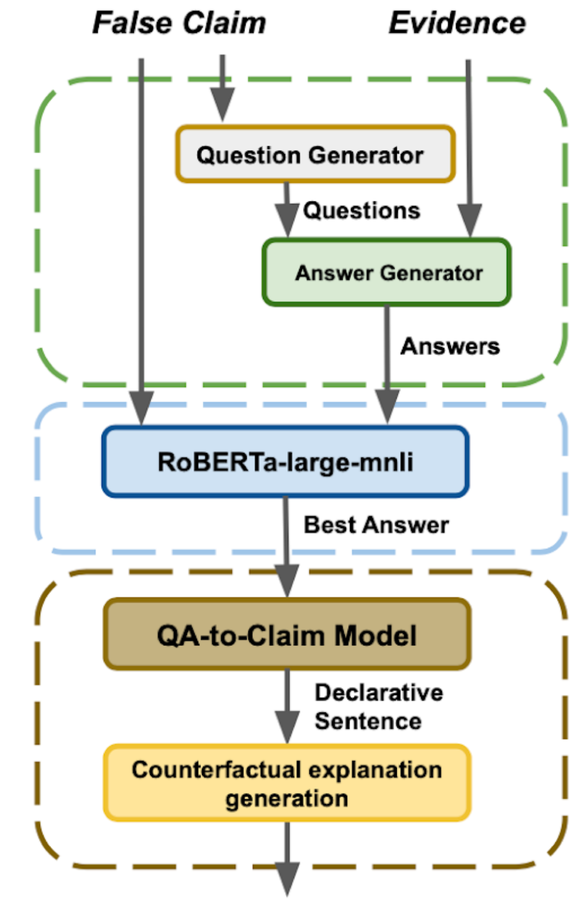
[1] Ruth M. J. Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19

[2] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. IJCAI-21

[3] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. ACL'21

[4] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. EMNLP'20

Methodology



Step 1. QA Generation

Step 2. Contradiction Checking

Step 3. Explanation Generation

We randomly selected 500 False Claims from the **FEVER** dataset to generate the CF explanations.

Error Analysis:

System error	Answer not correctly picked	25 (16.7%)
	Wrong grammar	9 (6%)
	Wrong answer/question	74 (43.7%)
Dataset error	Wrong claim label	6 (4%)
	Insufficient evidence	36 (24%)
Total error		150

Human Evaluation Result

We compared CF explanations with two SOTA summary-based model

- Extractive (EXT) : DistillBert [3]
- Abstractive (ABS): RoBERTa [4]

Both models were fine-tuned on CNN/Daily Mail dataset.

Survey 1: Compared the explainability of the three CF explanations for why a piece of news is fake. (425 participants. Each completed 5 samples)

Model	Best					Worst				
	Familiar (581)		Unfamiliar (407)		Overall (988)	Familiar (581)		Unfamiliar (407)		Overall (988)
	PR (513)	PF (68)	PR (120)	PF (287)		PR (513)	PF (68)	PR (120)	PF (287)	
CF-A	**0.41**	0.29	**0.42**	**0.40**	**0.40**	**0.31**	0.40	**0.31**	**0.27**	**0.31**
CF-N	0.32	**0.35**	0.28	0.33	0.32	0.34	**0.24**	0.37	0.33	0.34
CF-M	0.27	**0.35**	0.31	0.26	0.28	0.35	0.37	0.33	0.35	0.34

Proportion of each explanation being selected as the best or the worst explanation.

Survey 2: Compared the best CF explanation from Survey-1 with the SOTA summary-based methods. (625 participants. Each completed 3 samples)

Model	average ranking					average ranking*				
	Familiar (480)		Unfamiliar (485)		Overall (965)	Familiar (480)		Unfamiliar (485)		Overall (683)
	PR (416)	PF (64)	PR (136)	PF (349)		PR (416)	PF (64)	PR (136)	PF (349)	
CF-A	**1.86**	2.0	1.99	**1.78**	**1.86**	**1.92**	**1.87**	1.99	**1.72**	**1.86**
EXT	2.11	2.01	**1.97**	2.03	2.05	2.09	2.07	**1.92**	2.02	2.03
ABS	2.02	**1.98**	2.04	2.20	2.08	1.98	2.05	2.09	2.25	2.10

The average ranking * calculates the average ranking without any CF generation system errors.

Conclusion

CF method outperforms the existing SOTA summary-based methods by **0.19** ranking place

KDD 2022 August 14-18, 2022 Washington DC

Acknowledgement: This research are supported by Ministry of Science and Technology, Taiwan, under Grant no. 110-2221-E-001-001- and 110-2634-F-002- 051-. Further, this work is supported by Academia Sinica under AS IIS 3006-37C4218.
The works of Aiping Xiong were in part supported by NSF award 1915801.