Unsupervised Explanation Generation via Correct Instantiations

Sijie Cheng[1,2], Zhiyong Wu[1], Jiangjie Chen[2], Zhixing Li[3], Yang Liu[5], Lingpeng Kong[1,4]

[1]Shanghai Artificial Intelligence Laboratory

[2]Fudan University [3]Full Truck Alliance

[4]The University of Hong Kong [5]Tsinghua University

Email: sjcheng20@fudan.edu.cn

Explainable Natural Language Processing

Instance	Explanation
Premise: A white race dog wearing the number eight runs on the track. *Hypothesis:* A white race dog runs around his yard. *Label:* contradiction	(highlight) *Premise:* A white race dog wearing the number eight runs on the track. *Hypothesis:* A white race dog runs around his yard.
	(free-text) A race track is not usually in someone's yard.
Question: Who sang the theme song from Russia With Love? *Paragraph:* ...The theme song was composed by Lionel Bart of Oliver! fame and sung by Matt Monro... *Answer:* Matt Monro	(structured) *Sentence selection:* (not shown) *Referential equality:* "the theme song from russia with love" (from question) = "The theme song" (from paragraph) *Entailment:* X was composed by Lionel Bart of Oliver! fame and sung by ANSWER. ⊢ ANSWER sung X

Table 1: Examples of three explanation types.

Wiegreffe, S. and Marasović, A., 2021. Teach me to explain: A review of datasets for explainable nlp. *arXiv preprint arXiv:2102.12060*.

Free-text Explanation for False Statements

False Statement	Explanation	Conflict Point
John put an elephant into the fridge.	An elephant is much bigger than a fridge.	Volume
He drinks apple.	Apple can not be drunk.	Function
Jeff ran 100,000 miles today.	No way can someone run 100,000 miles in a day.	Speed
A giraffe is a person.	A giraffe is an animal, not human.	Property
Europe is in France.	Europe is a continent but france is a country.	Geography

Table 2: Examples and their exact conflict points to explain in ComVE task.

- Find the **Conflict Point** where the false statement contradicts the commonsense knowledge.

Wang, C.; Liang, S.; Jin, Y.; Wang, Y.; Zhu, X.; and Zhang, Y. 2020. SemEval-2020 Task 4: Commonsense Validation and Explanation. In SEMEVAL.

Challenges

- **(Supervision)** Manually constructing a dataset with conflict points for training is labor-intensive and difficult to scale.
- **(Explicit Knowledge)** Exact triples of conflict points are rare in the external knowledge graph due to their tacitness and diversity.

⇩

Inspired by the line of work about **the chain of thought**.

Provide **guided hints** as prompts to **implicitly** elicit Pre-trained Language Models (PLMs) to reason the conflict point automatically.

Framework

- **Phase1 (Correct Instantiations Generation) → Commonality.**
- **Phase2 (Explanation Generation) → Contrast.**

The PLMs can implicitly induce the conflict point better to generate explanations.

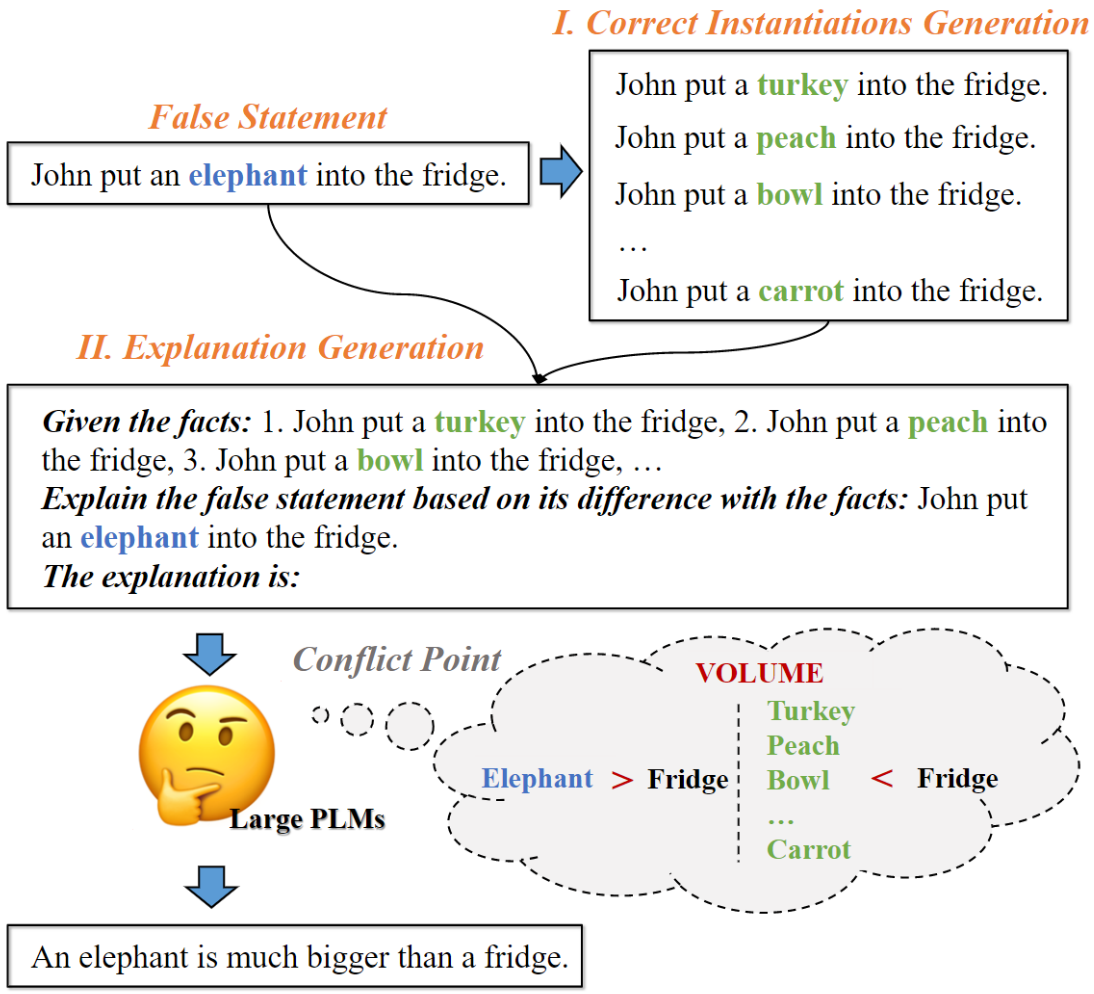

Figure 1: Our proposed two-phase framework NEON.

Phase1: Correct Instantiations Generation

- **In-context Learning** (Few-shot)

Task: Based on the incorrect statement, generate the correct statement.
/* Example 1 */
Incorrect statement: **He drinks apple.**
Correct statement: **He drinks milk.**
/* Test data */
Incorrect statement: **John put an elephant into the fridge.**
Correct statement:

Table 3: The prompt instances of in-context learning in the first phase.

- **Constrained Text Generation: CGMH** (Unsupervised)
  - Step 1: Where to Edit – Conflict Detection.

$$S^i_{\text{PPL}} = \frac{\text{PPL}(\boldsymbol{x})}{\text{PPL}(\boldsymbol{x} \backslash \{x^i\})}$$

  - Step 2: Edit with What – Modification Action.

$$S_{\text{Fluency}} = \prod_{i=1}^{n} P_{\text{LM}}(h^i | h^{<i})$$

Miao, N.; Zhou, H.; Mou, L.; Yan, R.; and Li, L. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 6834–6842.

Phase2: Unsupervised Explanation Generation

- **In-context Learning** (Zero-shot)
  - To purely detect the ability of implicit induction in off-the-shelf PLMs, we explore the model performance without any signals rather than supervised setup.

Given the facts: **1. John put a turkey into the fridge, 2. John put a peach into the fridge, 3. John put a bowl into the fridge,** Explain the following statement based on its difference with the facts: **John put an elephant into the fridge.** The explanation is:

Table 4: The prompt instances of in-context learning in the second phase.

Experiments

- **Model:** OPT-175B.
- **Datasets:** ComVE & e-SNLI.

Dataset	Preferred Explanation (%)			κ
	Original	Tie	NEON	
ComVE	20.33	42.67	37.00	0.47
e-SNLI	18.67	41.67	39.67	0.39
	Conflict Point (%)			
ComVE	19.33	46.00	34.67	0.45
e-SNLI	15.67	53.67	30.67	0.36

Table 5: The results of manual evaluation.

Method	ComVE				e-SNLI			
	BLEU	ROUGE	BERTScore	S-BERT	BLEU	ROUGE	BERTScore	S-BERT
Random	1.47	17.81	46.21	42.54	4.94	24.23	50.73	43.05
Retrieval-BM25	1.51	17.23	45.18	38.68	4.29	23.31	49.80	42.09
Retrieval-SBERT	1.69	18.55	46.64	45.47	4.64	24.45	51.16	48.22
Original	1.88	20.21	48.68	51.82	4.71	25.38	50.92	46.39
Ground-truth	2.48	21.25	49.66	**55.21**	5.57	25.62	51.96	49.19
Top-1	2.42	21.42	49.86	55.03	6.03	25.87	51.97	48.51
NEON w/ CGMH	3.37	20.10	48.92	49.50	4.67	26.04	51.04	48.42
NEON w/ In-context	**3.39**	**22.50**	**51.50**	54.52	**6.20**	**27.28**	**53.87**	**51.69**

Table 6: The results of automatic evaluation.

Analysis

- **Quality of Generated Instantiations**
  - **Automatic Evaluation:** fine-tune RoBERTa-Large on training datasets as binary classifiers with 88.97 and 84.25 accuracies.

Dataset	NEON	Human Generated
ComVE	70.28	89.60
e-SNLI	92.30	97.84

Table 7: The results of automatic evaluation.

  - **Manual Evaluation:** i. Acceptability; ii. Grammaticality; iii. Factuality; iv. Diversity; v. Commonality.

Dataset	Acc.	Gram.	Fact.	Diver.	Common.
ComVE	72.80	2.97	2.66	2.63	2.56
e-SNLI	81.67	2.88	2.72	2.89	2.66

Table 8: The results of manual evaluation.

Analysis

- **Effects on Instantiations Number.**

#	BLEU	ROUGE	BERTScore	S-BERT
1	2.42	21.03	49.22	52.70
2	2.61	21.14	49.22	52.56
3	3.32	21.32	49.46	51.79
4	3.29	22.26	50.97	**54.74**
5	3.39	**22.50**	**51.50**	54.52
6	3.01	21.49	49.11	49.06
7	**3.48**	21.57	49.45	49.66
8	3.28	21.27	49.66	49.94
9	3.16	21.70	49.91	48.73
10	3.39	21.21	49.94	49.47

Table 9: Model performance with increasing number of ensemble instantiations in the ComVE task.

- **Demonstration of Generality**
  - Generate explanation for correct statements in the e-SNLI task.
  - Directly use the generated correct instantiations as guided hints.

Method	BLEU	ROUGE	BERTScore	S-BERT
Original	8.11	29.73	52.66	53.18
Top-1	9.22	28.64	52.64	50.81
NEON	**11.18**	**31.69**	**55.30**	**56.33**

Table 10: Model performance of generating explanations for correct statements in the e-SNLI task.

Thanks!

Sijie Cheng