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

Introduction

Task: Free-text Explanation for False Statements

- **Task Definition:** given a false statement, the model is expected to generate a convincing free-text explanation to state the reason why the former statement is incorrect.
- **False Statement:** John put an elephant into the fridge.
- **Free-text Explanation:** An elephant is much bigger than a fridge.
- **Conflict Point:** Volume.

➢ **(The key point)** Find the **Conflict Point** where the false statement contradicts the commonsense knowledge.

Previous Studies and Limitations

- **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.

Motivation

➢ **(Solution)** Provide guided hints as prompts to implicitly elicit PLMs to reason the conflict point automatically, inspired by the line of work about the chain of thought.

Contribution

➢ We propose a novel method based on the importance of conflict points to solve the false statement explanation problem. To the best of our knowledge, we are the first to introduce the concept of the conflict point in the task.

➢ We propose a two-phase framework named Neon to elicit the large PLMs to induce through instantiations to unsupervised explanation generation.

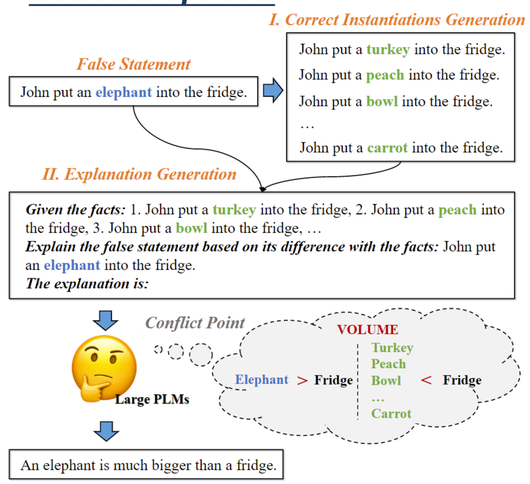
➢ We present analyses of our generated instantiations and demonstrate the generality of Neon.

The Neon Framework

Overview of Neon

Properties:

- Phase I: **Commonality**.
- Phase II: **Contrast**.



Phase I: 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:

- **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.

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

Phase2: Unsupervised Explanation Generation

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

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:

Experiments

Main Experiments

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

- **Manual Evaluation:**

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

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

- **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

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**

➢ **More analysis can be found in our paper.**