A Survey on Natural Language Counterfactual Generation

Yongjie Wang[1]*, Xiaoqi Qiu[2]*, Yue Yu[1], Xu Guo[1], Zhiwei Zeng[1], Yuhong Feng[2], Zhiqi Shen[1]

1.{yongjie.wang, yue.yu, xu.guo, zhiwei.zeng, zqshen}@ntu.edu.sg

2. qiuxiaoqi2022@email.szu.edu.cn, yuhongf@szu.edu.cn

1. Nanyang Technological University, Singapore. 2. Shenzhen University, China. *Equal Conribution

1 Introduction

Undesirable behaviors of LMs raise the demand for model explainability.

Counterfactual Generation can HELP ★

- create counterfactual examples (**CFEs**) with desired labels by minimal edits.
- highlight attributable factors to probe reasoning behind predictions ("what-if" scenarios).

Use cases of Counterfactual Generation

CFE in Sentiment Analysis Task : $(x, y) \rightarrow (c, y')$

This is a bad movie (Negative). ➡ This is a good movie (***Positive***).

CFE in Natural language Inference Task : $(x_p, x_h, y) \rightarrow (c_p, c_h, y')$

P: A child is creating sculptures.
H: A child is painting on canvas.
(Contradiction)

➡

P: A child is making something.
H: A child is painting on canvas.
(***Neutral***)

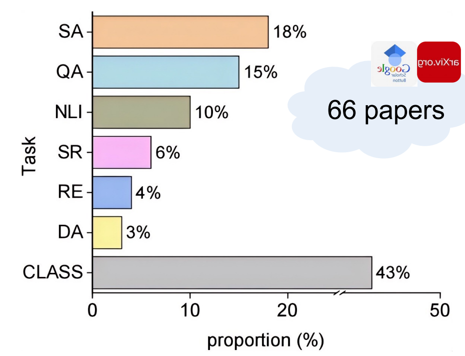
benefit ➡

Explanability
CFEs reflect model behaviors

Robustness
Counterfactually augment data

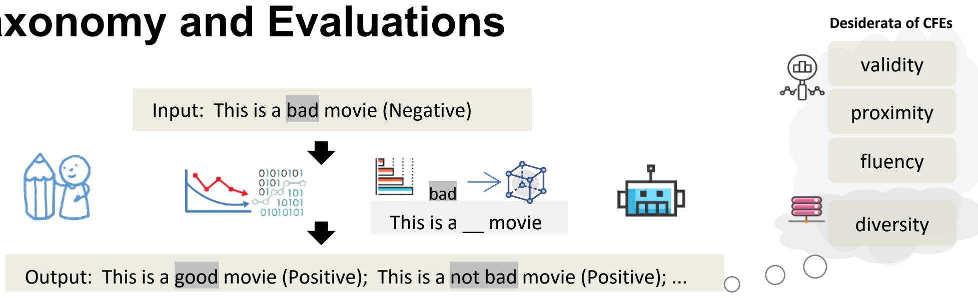
Fairness
CFEs help error analysis

2 Why this survey



- **Formulations** vary by specific tasks.
- Various **considerations**, e.g., diversity, proximity.
- Different implementations and solving strategies.

3 Taxonomy and Evaluations



	Manual Generation	Joint Learning-based Generation	Identify and then Generate	LLM prompting
Description	Instructing human annotators to revise a sentence	Training an end-to-end model that jointly minimizes the multiple objectives with user desiderata	Employing a divide-and-conquer strategy: identifying important words and then replacing them	Prompting LLMs as generators
Training	-	Yes	Optional	No
Pros	Meaningful and minimal revision, high quality	End-to-end, quantifiable objectives; easy to optimize the joint objective	Explainability; high controllability; precise edit	User-friendly; cheaper; no training
Cons	Time-consuming; labor-intensive; expensive	Hard to quantify each objective; trade-off over multiple objectives; lower controllability	Complicated workflow	Hard to tune prompts; rely on prompt quality

Human Evaluation ✚ Quantitative Evaluation

complementary

Property		Metric	Trend
Validity		Flip Rate	↑
Proximity	Lexical	BLEU (Papineni et al., 2002)	↑
		ROUGE (Lin, 2004)	↑
		METEOR (Denkowski and Lavie, 2011)	↑
		Levenshtein Dist. (Levenshtein et al., 1966)	↓
		Syntax Tree Dist. (Zhang and Shasha, 1989)	↓
	Semantic	MoverScore (Zhao et al., 2019)	↑
		USE Sim. (Cer et al., 2018)	↑
		SBERT Sim. (Reimers and Gurevych, 2019)	↑
Diversity	Lexical	Self-BLEU (Zhu et al., 2018)	↓
		Distinct-n (Li et al., 2016)	↑
		Levenshtein Dist. (Levenshtein et al., 1966)	↑
	Semantic	SBERT sim. (Reimers and Gurevych, 2019)	↓
		BERTScore (Zhang et al., 2020)	↓
Fluency		Likelihood Rate (Salazar et al., 2020)	(→ 1)
		Perplexity Score (Radford et al., 2019)	↓
Model Performance		Accuracy / F1-Score	↑
		Std of accuracy / F1-score in multiple runs	↓

4 Challenges and Future Directions

Fair evaluation

- No ground truth.
- The evaluations are conducted from incomparable angles. One method may excel in validity but lag in diversity.

Model privacy and security

- Higher exposure to attackers, e.g., model extraction risks.

Unlock LLM prompting

- Long-context CFEs generation
  - ➢ Quality of CFEs deteriorates with longer input sentences.
- Hard to improve CFE quality
  - ➢ why and how to design effective prompts remains unclear.
- Specific LLMs for CFEs
  - ➢ no fine-tuned LLMs for CFE generation.
- LLM hallucination
  - ➢ LLM may inject misleading content into CFE.
- Lower Controllability
  - ➢ hard to precisely control over changes.

5 Conclusion

To bridge the gap in understanding CFE generation in NLP, we

- ➢ propose a clear taxonomy of existing solutions and analyze pros and cons of methods in each groups;
- ➢ summarize the common evaluation metrics;
- ➢ highlight the research challenges, especially the untapped potential of LLMs in CFE generation.

NANYANG TECHNOLOGICAL UNIVERSITY SHENZHEN UNIVERSITY EMNLP 2024