Three Counterfactual Interpretations: Identification and Applications

Hannah Chen

2023/08/02

University of Virginia

Overview

Three Counterfactual Interpretations and Their Identification (Pearl, 1999)

Application: Spurious correlation in NLP

- Are All Spurious Features in Natural Language A like? An Analysis through a Causal Lens (Joshi, 2022)

Standard Counterfactual Definition

Causation: event E would not have occurred if it weren't for the cause C

For example, A particular exposure $\rightarrow$ Disease

"Probability that disease would not have occurred in the absence of exposure, given that disease and exposure did in fact occur"

It captures the notion of "**necessary** cause"

Three Counterfactual Intepretations

1) Necessary cause

2) Sufficient cause

3) Necessary and sufficient cause

Necessity and Sufficiency

Necessary condition

Air is necessary for human life

“John is unmarried” is necessary for “John is a bachelor”

“X is a rectangle” is necessary for “X is a square”

Sufficient condition

Lighting is sufficient for thunder

“John is king” is sufficient to know “John is a male”

“X is a square” is sufficient for “X is a rectangle”

Necessity and Sufficiency (Wiki)
The Concepts of Necessary Conditions and Sufficient Conditions (Norman Swartz)

Necessity and Sufficiency

Propositional Logic

If X then Y ($X \rightarrow Y$),

1) Y is necessary for X
2) X is sufficient for Y

Logically converse

Causal Explanation

X is a necessary cause of Y

$\not\leftrightarrow$ Y is a sufficient cause of X

For example:

Lighting is a **sufficient condition** for thunder

Thunder is a **necessary condition** for lightning

$\rightarrow$ Not causal, $\because$ lightning causes thunder

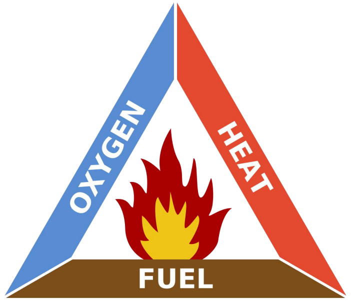
Lighting is a **sufficient cause** for thunder

Thunder is **not** a **necessary cause** for lighting

Why do we care?

Necessary causation → various factors would qualify as explanations

Oxygen → fire



Singular-event considerations

Necessary but not sufficient

Sufficient causation → we lose important specific information

Skipping the final exam → failing the course

Generic tendencies

Sufficient but not necessary

Other causes: poor attendance, procrastination, teaching style

The distinction between the two is imporant, especially when generating explanations for AI systems.

Definitions

Notation: Let X and Y be two binary variables in a causal model

$$x : X = \text{true}, y : Y = \text{true}$$

$$x' : X = \text{false}, y' : Y = \text{false}$$

U: exogenous,
V: endogenous, including X and Y

$$Y_x(u) : \text{ potential response of } Y \text{ to } do(X = x)$$

Necessary and Sufficient Cause

	Factual world		Counterfactual world	
Necessity	$X = \text{true}, Y = \text{true}$		$X = \text{false}, Y = \text{false}$	Does the absence of X still capable of producing Y?
	Exposed, has the disease	→	Unexposed, has no disease	
Sufficiency	$X = \text{false}, Y = \text{false}$		$X = \text{true}, Y = \text{true}$	Does the existence of X produces Y?
	Unexposed, has no disease	→	Exposed, has the disease	

Generally, not identifiable

Definitions

Probability of necessity (PN)

Probability that y would not have occured in the absence of x given that x and y did in fact occur

$$\text{PN} \triangleq P(Y_{x'} = \text{false} \mid X = \text{true}, Y = \text{true}) \triangleq P(y'_{x'} \mid x, y)$$

Probability of sufficiency (PS)

Probability that setting x would produce y given that x and y are in fact absent

$$\text{PS} \triangleq P(Y_x = \text{true} \mid X = \text{false}, Y = \text{false}) \triangleq P(y_x \mid x', y')$$

Probability of necessity and sufficiency (PNS)

Probability that y would respond to x in both ways

$$\text{PNS} \triangleq P(y_x, y'_{x'})$$

Example: Betting against a Fair Coin Toss

x: “bet on heads”, y: “win a dollar”, u: “the coin turned up heads”

Q: Was the bet a necessary cause (sufficient cause, or both) for winning?

Functional relationship: $y = (x \wedge u) \vee (x' \wedge u')$

$$\text{PN} = P(y'_{x'} \mid x, y) = P(y'_{x'} \mid u) = 1 \qquad \because x \wedge y \Rightarrow u \text{ and } Y_{x'}(u) = \text{false}$$

$$\text{PS} = P(y_x \mid x', y') = P(y_x \mid u) = 1 \qquad \because x' \wedge y' \Rightarrow u \text{ and } Y_x(u) = \text{true}$$

$$\begin{aligned}\text{PNS} &= P(y_x, y'_{x'}) \\ &= P(y_x, y'_{x'} \mid u)P(u) + P(y_x, y'_{x'} \mid u')P(u') \\ &= 1\frac{1}{2} + 0\frac{1}{2} = \frac{1}{2}.\end{aligned}$$

Example: Betting against a Fair Coin

x: “bet on heads”, y: “win a dollar”, u: “the coin turned up heads”

Q: Was the bet a necessary cause (sufficient cause, or both) for winning?

Functional relationship: $y = (x \wedge u) \vee (x' \wedge u')$ → To compute counterfactuals we need to know this

$$\mathrm{PN} = P(y'_{x'} \mid x, y) = P(y'_{x'} \mid u) = 1$$

→ The bet was 100% necessary for the win

$$\mathrm{PS} = P(y_x \mid x', y') = P(y_x \mid u) = 1$$

→ The bet was 100% sufficient for the win

$$\begin{aligned}\mathrm{PNS} &= P(y_x, y'_{x'}) \\ &= P(y_x, y'_{x'} \mid u)P(u) + P(y_x, y'_{x'} \mid u')P(u') \\ &= 1\frac{1}{2} + 0\frac{1}{2} = \frac{1}{2}.\end{aligned}$$

Betting heads has 50% chance of being necessary and sufficient cause of winning

Spurious Features in NLP

Spurious features: undesirable feature-label correlation, features model should not rely on

(Joshi, 2022): Features can be spurious for different reasons

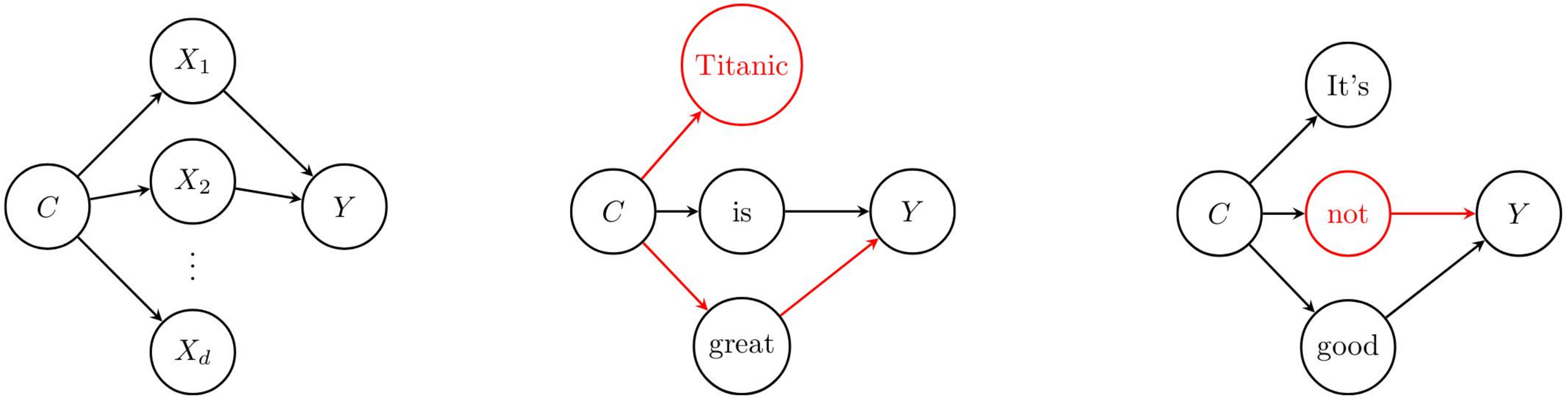
Irrelevant features
Speilberg's new film is brilliant. ⟶ Positive
______'s new film is brilliant. ⟶ Positive
Necessary features
The differential compounds to a hefty sum over time.
The differential will not grow ⟶ Contradiction
The differential will ___ grow ⟶ ?

Table 1: Difference between two spurious features: (a) the director name can be replaced without affecting the sentiment prediction; (b) the negation word is necessary as it is not possible to determine the label without it.

Most work focus on necessary but not sufficient spurious features

Causal Models for Text Classification

$X=(X_1,X_2,\ldots X_n)$: sequence of input words/features
Y: sentiment label
C: common cause of the input



(a) Data generating model. (b) Type 1 dependence. (c) Type 2 dependence.

<u>Non-causal association</u>

"Titanic" and Y are dependent because of confounder C

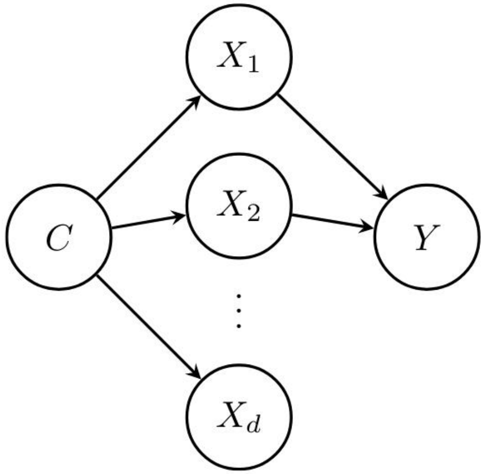
<u>Causal association</u>

"not" and Y are dependent

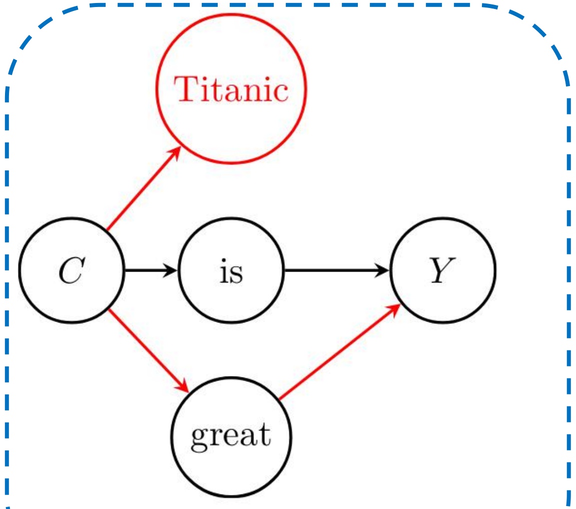
Causal Models for Text Classification

$X=(X_1,X_2,\ldots X_n)$: sequence of input words/features
Y: sentiment label
C: common cause of the input

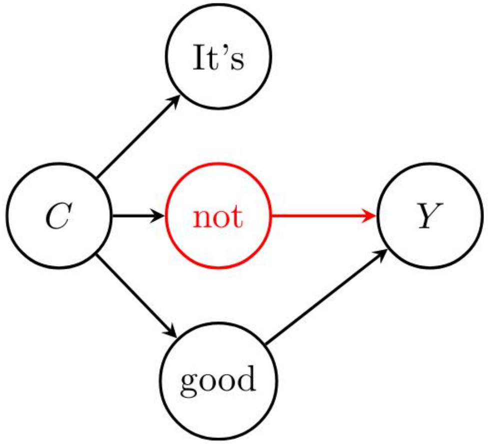
Spurious correlation



(a) Data generating model.



(b) Type 1 dependence.



(c) Type 2 dependence.

<u>Non-causal association</u>

"Titanic" and Y are dependent because of confounder C

<u>Causal association</u>

"not" and Y are dependent

Estimating PN & PS of a Feature

X_{-i}: context without X_i

PN, PS are *context dependent*

PN: probability that y would change if feature X_i were set to a different value

$$PN(X_i{=}x_i, Y{=}y \mid X_{-i}{=}x_{-i}) \triangleq p(Y(\underline{X_i \neq x_i}) \neq y \mid X_i{=}x_i, X_{-i}{=}x_{-i}, Y{=}y)$$

Intervention: text infilling with masked LMs, e.g., Titanic → Ip Man

PS: probability that setting X_i to x_i would produce y given x_i is absent

$$PS(X_i{=}x_i, Y{=}y \mid X_{-i}{=}x_{-i}) \triangleq p(Y(\underline{X_i{=}x_i}) = y \mid X_i{\neq}x_i, X_{-i}{=}x_{-i}, Y{\neq}y)$$

Intervention, e.g., adding negation

Estimating PN & PS of a Feature

Average effect of a feature: marginalize over the contexts

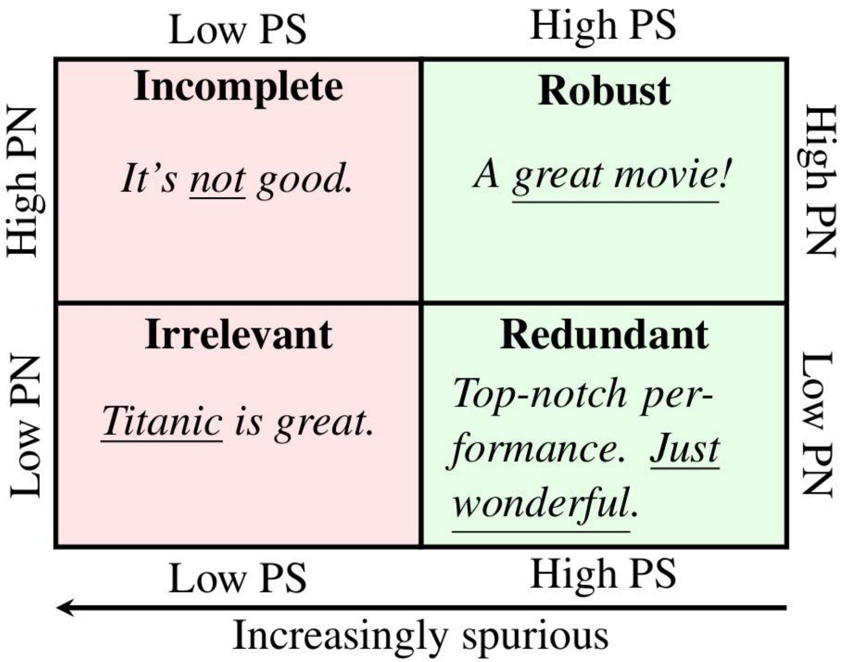
$$\mathrm{PN}(x_i, y) \triangleq \int \mathrm{PN}(x_i, y \mid X_{-i}) p(X_{-i} \mid x_i, y) \, \mathrm{d}X_{-i}$$

Spuriousness of a feature: $1 - PS(x_i, y)$

Spurious feature: if spuriousness > 0

Non-spurious feature: if sufficient in any context (high PS)

Feature Categorization



Calculating PN & PS requires knowing how the label would change when removing or adding a feature

P: The doctor was paid by the actor.	
H0: The actor paid the doctor.	**L0**: Entailment
H1: The teacher paid the doctor.	**L1**: Neutral
H2: The actor liked the doctor.	**L2**: Neutral
H3: The actor paid the guard.	**L3**: Neutral
H4: An actor paid the doctor.	**L4**: Entailment

High word overlap has high PN (but low PS) to entailment

Changing overlapped words is likely to change the label
Unless replaced with a synonym

Implications on Model Robustness

Is relying on spurious features always bad?

Prior work suggested models shouldn't rely on a single feature in any way

Model prediction should depend on high PN spurious features

It's only bad when model over relies on them and ignores other necessary features

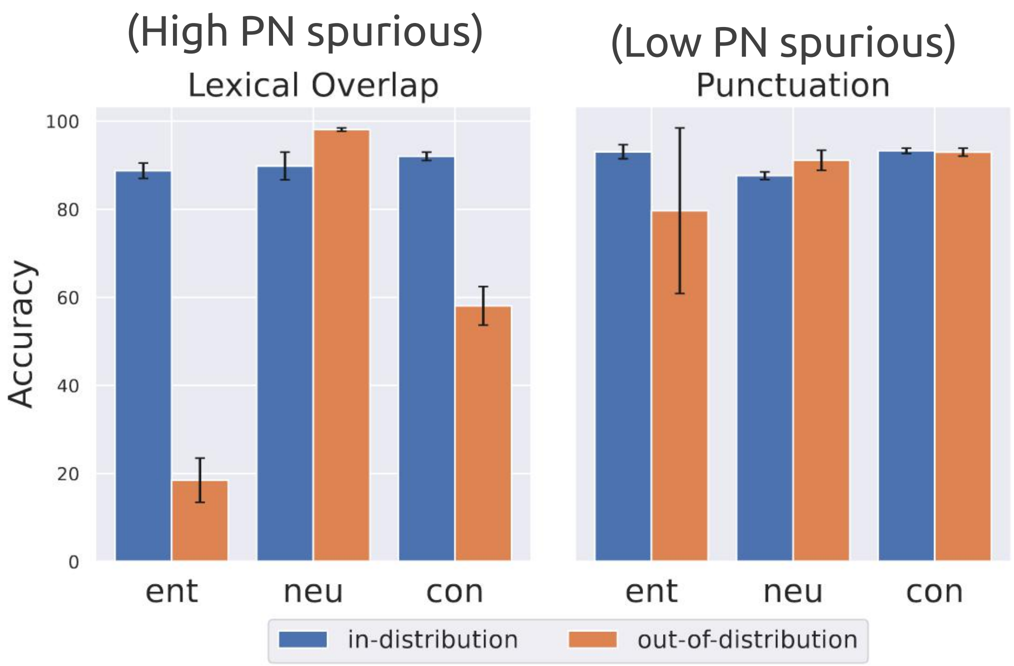
How to evaluate model robustness?

Common way: perturb and see if prediction is invariant → only tells us if the feature is necessary

This only works on testing robustness to low PN spurious features

Robustness to high PN spurious features: create test examples with same spurious feature but different label (e.g., HANS dataset, label flipping adversarial attacks)

Implications on Learning Methods

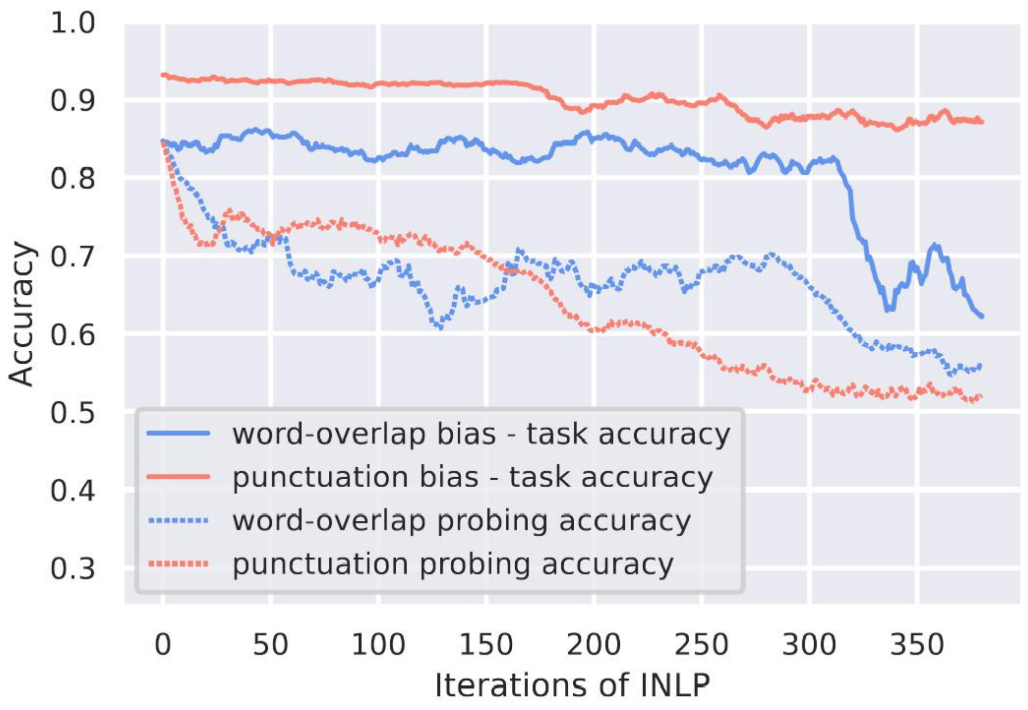


Debiasing via subsampling

Lexical Overlap: high overlap (in-distribution)
low overlap (OOD)

Punctuation ("!!"): with punctuations (in-distribution)
without punctuations (OOD)

Implications on Learning Methods



Debiasing via INLP (Ravgogel, 2020)

Removing spurious feature from learned representations

High PN spurious features: harder to remove, and hurts task performance