Causal Direction of Data Collection Matters: Implications of **Causal and Anticausal Learning for NLP**

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS
ETH zürich
UNIVERSITY OF CAMBRIDGE
UCL
IIT Kharagpur

Zhijing Jin*, Julius von Kügelgen*, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, Bernhard Schoelkopf

arxiv.org/abs/2110.03618

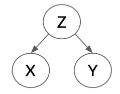
Concept of "Causality"

We observe data in the world

X	Y	Z	...
...	...	...	...
...	...	...	...

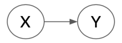
We want to infer the causal graph (e.g., "how were the data generated")?

By scientific discovery:



Example: Z causes both X and Y

By recovering the data generating process**:**



Example: First, X is independently generated. Following X, *Y is generated given the value of X.*

"Causal Direction" in Data Collection

[German Original EuroParl Transcript]
Wir haben es **heute** mit unbestrittenen Forderungen zu tun.

The translator chose to put "today" to the beginning of the sentence.
Alternative: Put "today" to the middle or the end.

[English Translation]
Today, it is undisputed claims that we are discussing.

Causal vs. Anticausal Learning (Schölkopf et al., 2012)

Data Collection Stage

First, we have	[German Original Transcript]
Then, we get	[English Translation]

P(translated_En | original_De) contains
- The habits of English translators
- General linguistic phenomena of German-to-English translationese

Modeling Stage

P(translated_En | original_De)	**Causal Learning** — *Same as* the data generation direction
P(original_De | translated_En)	**Anticausal Learning** — *Opposite to* the data generation direction

On causal and anticausal learning (Schölkopf et al., 2012).

Categorization of NLP Tasks

Bidirectional NLP Tasks

Data Collection Stage

	Translation		Questions and Answers	
First, we have	[Original German]	[Original English]	[SQuAD Answers]	[Questions from Natural Questions]
Then, we get	[Translated English]	[Translated German]	[SQuAD Questions]	[Answers from Natural Questions]

(Rajpurkar et al., 2016) (Kwiatkowski et al., 2019)

*Other bidirectional tasks: text style transfer, intent classification, etc.

Modeling Stage

Causal Learning — *Same as* the data generating process

Anticausal Learning — *Opposite to* the data generating process

Causal Learning Only

Data Collection Stage

	Summarization	Parsing	Tagging	IE	WebNLG
First, we have	Original Article	Text	Text	Text	KG triples
Then, we get	Summary	Syntax Tree	Part-of-Speech Tags	Structured Info (named entities, relations, etc.)	Crowdsourced Text

(Gardent et al., 2017)

Modeling Stage

Causal Learning — *Same as* the data generating process

Anticausal Learning Only

Data Collection Stage

	Author Attribute Classification	Review Sentiment Classification
First, we have	Author Attributes	Sentiment/Rating
Then, we get	Text	Review

Modeling Stage

Anticausal Learning — *Opposite to* the data generating process

Top 3 Messages of the Paper

Does the task nature (causal vs. anticausal) matter for ML?

Top 3 Messages of Our Paper:

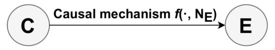
1. **Yes, it matters** (Sec 3 & 4)
2. **How does it matter?**
   => In semi-supervised learning and domain adaptation (Sec 5 & 6)
3. **What does this fact mean for future NLP practices?**
   => Causality-aware data collection & modeling (Sec 7)

Independent Causal Mechanism (ICM)

Data Collection Stage

First, we have	Variable C (Cause)
Then, we get	Variable E (Effect)

N_E (Annotation Preferences/Noise)



Modeling Stage

Causal Learning — *Same as* the data generating process: Input $\xrightarrow{f\text{ : Given C, predict E}}$ Output

Anticausal Learning — *Opposite to* the data generating process: Output $\xleftarrow{g\text{ : Given E, predict C}}$ Input

ICM Principle

The causal mechanism $f(\cdot, N_E)$ is independent of the cause C.

Algorithmic information (by Kolmogorov Complexity K): (Solomonoff, 1964; Kolmogorov, 1965)

$$K(P_{C,E}) \stackrel{+}{=} \boxed{K(P_C)} + \boxed{K(P_{E|C})} \quad \text{Decomposition (with no shared info)}$$
$$\stackrel{+}{\leq} K(P_E) + K(P_{C|E})\,.$$

(Janzing and Schölkopf, 2010)

The previous inequality of Kolmogorov complexity can be approximated by

$$\mathrm{MDL}(\mathbf{c}_{1:n}, \mathbf{e}_{1:n}) = \mathrm{MDL}(\mathbf{c}_{1:n}) + \mathrm{MDL}(\mathbf{e}_{1:n}|\mathbf{c}_{1:n})$$
$$\leq \mathrm{MDL}(\mathbf{e}_{1:n}) + \mathrm{MDL}(\mathbf{c}_{1:n}|\mathbf{e}_{1:n}), \qquad (2)$$

Experiment on MT data confirms the inequality

Dataset	Size	Note
En→Es	81K	Original English, Translated Spanish
Es→En	81K	Original Spanish, Translated English
En→Fr	16K	Original English, Translated French
Fr→En	16K	Original French, Translated English
Es→Fr	15K	Original Spanish, Translated French
Fr→Es	15K	Original French, Translated Spanish

Table 2: Details of the CausalMT corpus.

Implications of ICM

Takeaway #1:

Alignment with the data generating process matters
=> It will make the learning task causal vs. anticausal

Takeaway #2:

Semi-supervised learning: Anticausal learning **shows** +1.70 % average improvement confirmed by a meta-study across 55 causal and 50 anticausal tasks, as opposed to +0.04% for causal learning.

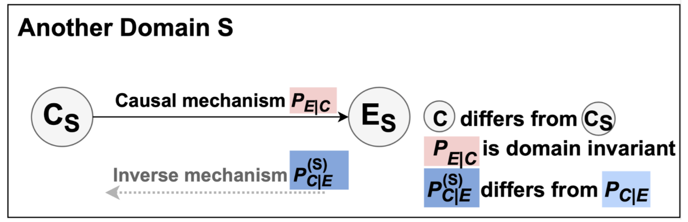
Domain adaptation: Causal learning **shows** +5.18 % average improvement confirmed by a meta-study across 22 causal and 11 anticausal tasks, as opposed to +1.26% for anticausal learning.

Suggested future work: *Causality-aware data collection*
Causality-aware modeling
E.g., different modeling designs for causal and anticausal learning

Our follow-up work on causality+MT will come out soon.

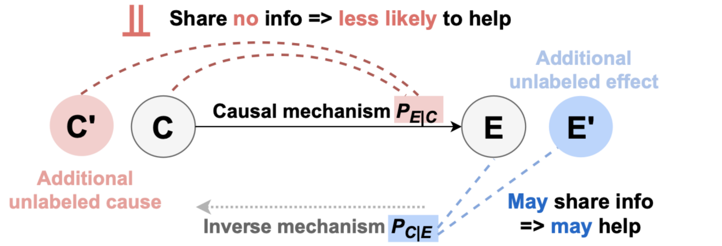
ICM's Implications for Domain Adaptation (Schölkopf et al., 2012)

Domain adaptation with covariate shift



ICM's Implications for Semi-Supervised Learning

(Schölkopf et al., 2012)



Causality-Aware Data Collection & Modeling

- When **collecting data,**
  Also annotate the **causal direction** between the input and output.
  E.g., [Cause] Input or [Effect] Input
  [Effect] Output [Cause] Output
- When **modeling,** consider the causal direction as an important piece of info
  - E.g., add tag [Cause] or [Effect] to the input and output data
  - E.g., add a task prefix "[Modeling-Effect-to-Cause]" (to trigger different inference behavior) (Riley et al., 2020)
  - E.g., limiting self-training (as an SSL method) only to the anticausal direction (Shen et al., 2021)

References:
1. On causal and anticausal learning (Schölkopf et al., 2012)
2. Causal inference using the algorithmic Markov condition (Janzing and Schölkopf, 2010)
3. Elements of causal inference: foundations and learning algorithms(Peters et al., 2017)
4. Mining the Cause of Political Decision-Making from Social Media: A Case Study of COVID-19 Policies across the US States (Jin et al., 2021)