CausalNLP Tutorial: An Introduction to Causality for Natural Language Processing

Zhijing Jin
Max Planck Institute & ETH Zürich
jinzhi@ethz.ch

Amir Feder
Columbia University & Google Research
amir.feder@columbia.edu

Kun Zhang
Carnegie Mellon University & MBZUAI
kunz1@cmu.edu

Abstract

Causal inference is becoming an increasingly important topic in deep learning, with the potential to help with critical deep learning problems such as model robustness, interpretability, and fairness. In addition, causality is naturally widely used in various disciplines of science, to discover causal relationships among variables and estimate causal effects of interest. In this tutorial, we introduce the fundamentals of causal discovery and causal effect estimation to the natural language processing (NLP) audience, provide an overview of causal perspectives to NLP problems, and aim to inspire novel approaches to NLP further. This tutorial is inclusive to a variety of audiences and is expected to facilitate the community's developments in formulating and addressing new, important NLP problems in light of emerging causal principles and methodologies.

1 Introduction

Establishing causal relationships is a fundamental goal of scientific research (Pearl, 2000; Spirtes et al., 2001). It naturally boils down to questions of causality when we need to quantify the effectiveness of a vaccine, the persuasive power of a public health ad, or the impact of a lockdown policy: How would the treatment (e.g., vaccine) affect the outcome (e.g., infection rates) compared to a counterfactual world with no treatment? Once formally identified, the direction and strength of causal relationships play a key role in the formulation of clinical treatments, public policy, and other long-standing prescriptive strategies. With such broad applications, a growing body of literature focuses on the interplay between NLP and causal inference (Tan et al., 2014; Wood-Doughty et al., 2018; Sridhar and Getoor, 2019; Veitch et al., 2020; Keith et al., 2020; Feder et al., 2021c).

Despite the interdisciplinary interest in causal inference with text, research in this space seems to remain scattered across domains without clear definitions, notations, benchmark datasets, and an understanding of the state of the art and challenges that remain. For example, it is unclear how deficiencies in NLP methods (such as their inaccuracy with low-resource languages and their tendency to propagate biases in the data) affect downstream causal estimates. In addition, hyperparameter selection and modeling assumptions in NLP are motivated by accuracy and tractability considerations; how these choices affect downstream causal estimates is underexplored.

This tutorial aims to address three questions: (1) What is causality? (2) How can the causal formulation help improve NLP models? (3) How can causality help NLP and computational social science to discover causal phenomena in our society?

Specifically, we will first introduce the fundamentals of causality for the NLP audience, then review research using the causal formulation to help NLP models (in terms of robustness, fairness, and interpretability), and finally introduce how causality can help NLP and computational social science to discover causal relations behind social phenomena.

2 Tutorial Overview

This introductory tutorial aims to introduce causality to the NLP research community. While causality plays a major role in scientific research, it has only now started to disseminate into the NLP community. This is why this tutorial will first focus on providing a generalized introduction to causality and its importance and relevance to the NLP community. We will then dive into the intersection of causality and NLP, and divide it into two distinct areas: using causal formalisms to make NLP methods more interpretable, robust and fair, and discovering causal relations in social phenomena that involve text variables. Accordingly, we divide the content of the tutorial into the following three parts:

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 17 - 22

1. Introduction to Causality. We will give a broad coverage of central concepts, principles, and technical developments in causal modeling; identification of causal effects (known as causal effect estimation); and finding causal relations by analyzing observational data (known as causal discovery). We will focus on representations and usage of causal models (Pearl, 2000; Spirtes et al., 2001), how causality is different from and connected to association, and recent machine learning methods for causal discovery and causal representation learning (Spirtes et al., 2001; Peters et al., 2017; Spirtes and Zhang, 2016; Shimizu et al., 2006; Zhang and Hyvärinen, 2009; Xie et al., 2020, 2022; Huang et al., 2022; Yao et al., 2022).

Specifically, we will answer the following questions: How can we define causality? Is causality an indispensable notion in science and machine learning? Why do we care about causality? How can we infer the causal effect of one variable on another? How can one learn causality from purely observational data? How can we recover latent causal variables and their relations?

2. Causality to Help Improve NLP Models. In this part of the tutorial, we will first motivate the audience by introducing why and how the causal perspective helps in a class of machine learning or AI tasks (Schölkopf et al., 2021; Pearl and Bareinboim, 2011; Schölkopf et al., 2012; Zhang et al., 2013, 2020). Briefly, although deep learning models achieve impressive performance by using correlations for prediction tasks, there are still limitations in their robustness, interpretability and fairness, which could be improved using causality.

With these motivations, we will then extend the causal formulation to NLP. Here, we will identify and highlight existing limitations in NLP methods, and will propose three application areas where causal ideas might help: interpretability (Guidotti et al., 2018), robustness (e.g., McCoy et al., 2019) and fairness (e.g., Zhao et al., 2017). For each potential application area, we will highlight the relevance of causal thinking in solving important open problems in NLP (Feder et al., 2021c; Veitch et al., 2021; Kilbertus et al., 2017).

3. Causality for NLP and Computational Social Science. Distinct from how causality can help improve NLP models in Part 2, we can also see another important use of NLP: identifying causal relations for NLP and computational social science. For example, does there exist gender bias in the upvotes of online posts (Veitch et al., 2020)? Do social media opinions affect the strictness of the COVID-19 social distancing policies (Jin et al., 2021b)? What are the reasons behind popular tweets? Many of these social problems involve text data. For example, online posts, news articles, scientific papers, conversation records, and many others are all text variables. If we want to investigate causal questions, such as the effect of certain contents or features of text on a certain outcome, then we need to run statistical causal models with text modeling.

In this part, we will first introduce how to conduct text-involved causal effect estimation discovery and causal discovery. Then, we will cover some real-world examples where we can apply these methods (Veitch et al., 2020; Feder et al., 2021b; Jin et al., 2021b; Ding et al., 2022; Keidar et al., 2022), and finally run through some exercise questions.

3 Tutorial Outline

For the three-hour tutorial, we will use 2.5 hours to cover three main topics: introduction of causality, how causality can help improve NLP models, and how causality can be applied to NLP and computational social science. And finally, we will use the remaining 30 minutes for an interactive exercise and Q&A.

An outline of the tutorial content is as follows:

1. Introduction to causality (60-min lecture + 5-min break)
   - Motivations: What is causality? Why is causality helpful for NLP?
   - Main content: Basics of causal effect estimation, causal discovery, and causal representation learning.
   - Example work: Pearl (2000); Feder et al. (2021b); Xie et al. (2020); Yao et al. (2022).
2. Causality to help improve NLP models (60-min lecture + 5-min break)
   - Motivations: If the goal is to help improve NLP models, how can causality help? What are some use case examples?
   - Main content: Inspirations from causality to help a variety of NLP topics: model robustness, domain adaptation, debiasing models, interpretability, and fairness.
   - Example work: Schölkopf et al. (2021); Feder et al. (2021c); Veitch et al. (2021);

Jin et al. (2021c); Stolfo et al. (2022); Hupkes et al. (2022).

3. Applications of causality for NLP and computational social science (20-min lecture)
   - Motivations: If the goal is to identify causal phenomena in our society, how can we learn causality on variables that involve text?
   - Main content: Use of SCMs and potential outcomes for NLP social applications such as explaining social media behavior, political phenomena, effective education, and gender bias in the research community. We will also cover cases where causal discovery and inference can help verify linguistic theories.
   - Example work: Veitch et al. (2020); Jin et al. (2021b); Ding et al. (2022).
4. Interactive exercise (20 min)
   - Given a social application of NLP, we will let the audience draw the causal graph, and brainstorm interesting research questions.
   - Given a fairness question in NLP, we will let the audience draw the causal graph, and discuss the causal formulation.
5. Q&A (10 min)

4 Tutorial Breadth

As for the contents of this tutorial, we will mainly cover beginner-friendly introductory materials of NLP, from the studies of established causality researchers out of the NLP domain, such as Judea Pearl, Donald Rubin, Bernhard Schölkopf, Clark Glymour, and Peter Spirtes. Apart from the work from these causality researchers, when it comes to the more specific connection of NLP and causality, we will cover the research work of various researchers: Dyanya Sridhar (Mila), Victor Veitch (University of Chicago), Zach Wood-Doughty (Northwestern University), Justin Grimmer (Stanford), Brandon M. Stewart (Princeton), Margaret E. Roberts (UCSD), Reid Pryzant (Microsoft), and many others.

5 Organizing Committee

Zhijing Jin (she/her) is a PhD at Max Planck Institute and ETH Zürich supervised by Prof Bernhard Schölkopf. Her research aims to (1) improve NLP models by connecting NLP with causal inference (Jin et al., 2021c,b; Ni et al., 2022), and (2) expand the impact of NLP by promoting NLP for social good (Jin et al., 2021a; Field et al., 2021; Gonzalez et al., 2022). She has published at many NLP and AI venues (e.g., AAAI, ACL, EMNLP, NAACL, COLING, AISTATS), and NLP for healthcare venues (e.g., AAHPM, JPSM). To foster the causality research community, she serves as the Publications Chair for the 1st conference on Causal Learning and Reasoning (CLeaR) (Schölkopf et al., 2022). She is also actively involved in AI for social good, as the organizer of NLP for Positive Impact Workshop at ACL 2021 (Field et al., 2021) and EMNLP 2022, and RobustML workshop at ICLR 2021. To support the NLP research community, she organizes the ACL Year-Round Mentorship program from 2021.

Amir Feder (he/him) is a postdoc at Columbia University, working with Prof David Blei. Amir develops methods that integrate causality into natural language processing to generate more robust and interpretable models. He is also interested in investigating and developing linguistically informed algorithms for predicting and understanding human behavior. Amir is currently also a visiting researcher (part time) at Google Research's Medical Brain Team, where he works on methods that leverage causal methodology for medical language models. He is a co-organizer of the First Workshop on Causal Inference and NLP (CI+NLP) at EMNLP 2021 (Feder et al., 2021a).

Kun Zhang (he/him) is an associate professor at Carnegie Mellon University and MBZUAI. His research interests lie in causal discovery and causality-based learning. He develops methods for automated causal discovery from various kinds of data, investigates learning problems including transfer learning and deep learning from a causal view, and studies philosophical foundations of causation and machine learning. He co-authored a best student paper for the Conference on Uncertainty in Artificial Intelligence (UAI) and a best finalist paper for the Conference on Computer Vision and Pattern Recognition (CVPR), and received the best benchmark award of the 2nd causality challenge. He has taken essential roles at many events of causal inference, including the general and program co-chair of the 1st Conference on Causal Learning and Reasoning (CLeaR 2022), program co-chair of the UAI 2022, co-organizer of the 9th Causal Inference Workshop at UAI 2021, co-organizer of NeurIPS 2020 Workshop on Causal Discovery and Causality-Inspired Machine Learn-

ing, 2020, co-editor of a number of journal special issues on causality, and many others.

6 Diversity Efforts

Our organizing committee includes both junior and senior instructors, as well as diverse genders, racial/ethnic backgrounds, and affiliations across America, Europe and Asia, which will help make people from various backgrounds feel more welcome to our workshop.

The topic of our workshop is causal inference, which can serve as a helpful tool for many NLP tasks, and the methods can scale up to various languages and domains. In addition, we advertise the tutorial to diversity-oriented venues (e.g., Widening NLP, QueerInAI, BlackInAI, WiML).

7 Target Audience & Prerequisites

There is no required audience background. Preferred knowledge includes the basics of statistics (e.g., understanding of probability distribution of single variables, joint probability distributions, and conditional probability distributions), and the basics of NLP (e.g., understanding of sentence embeddings, and the setup of simple NLP tasks such as classification).

8 Recommended Reading List

We compiled a recommended reading list of causality and NLP papers at (Jin, 2021).[1] Among the papers, the top three recommended readings are Guo et al. (2020), Schölkopf et al. (2021) and Feder et al. (2021b).

9 Other Information

Tutorial Type: Introductory.
Tutorial Materials: We will make available on our GitHub (Jin, 2021) all tutorial presentation materials, including slides, captioned video recordings, codes, and the recommended paper list.

10 Ethical Considerations

The theme of the tutorial focuses on introducing the method of causal inference to NLP. The introduction materials will stay on the technical side. There will not be direct links to applications that will raise ethical concerns. Additionally, since one of the instructor's research background is NLP for social good, we will introduce some use cases of NLP and causal inference for social good applications.

[1] https://github.com/zhijing-jin/Causality4NLP_Papers

References

Yiwen Ding, Jiarui Liu, Zhiheng Lyu, Kun Zhang, Bernhard Schoelkopf, Zhijing Jin, and Rada Mihalcea. 2022. Voices of her: Analyzing gender differences in the AI publication world.

Amir Feder, Katherine Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Molly Roberts, Uri Shalit, Brandon Stewart, Victor Veitch, and Diyi Yang, editors. 2021a. *Proceedings of the First Workshop on Causal Inference and NLP*. Association for Computational Linguistics, Punta Cana, Dominican Republic.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2021b. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *CoRR*, abs/2109.00725.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021c. CausaLM: Causal model explanation through counterfactual language models. *Comput. Linguistics*, 47(2):333–386.

Anjalie Field, Shrimai Prabhumoye, Maarten Sap, Zhijing Jin, Jieyu Zhao, and Chris Brockett, editors. 2021. *Proceedings of the 1st Workshop on NLP for Positive Impact*. Association for Computational Linguistics, Online.

Fernando Gonzalez, Zhijing Jin, Jad Beydoun, Bernhard Schölkopf, Tom Hope, Mrinmaya Sachan, and Rada Mihalcea. 2022. How is NLP addressing the UN Sustainable Development Goals? a challenge set to analyze NLP for social good papers.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Comput. Surv.*, 53(4):75:1–75:37.

Biwei Huang, Charles Low, Feng Xie, Clark Glymour, and Kun Zhang. 2022. Latent hierarchical causal structure discovery with rank constraints. In *Neural Information Processing Systems (NeurIPS)*.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian

Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2022. State-of-the-art generalisation research in NLP: a taxonomy and review. *CoRR*, abs/2210.03050.

Zhijing Jin. 2021. Causality for NLP reading list. https://github.com/zhijing-jin/Causality4NLP_Papers.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021a. How good is NLP? A sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics.

Zhijing Jin, Zeyu Peng, Tejas Vaidhya, Bernhard Schoelkopf, and Rada Mihalcea. 2021b. Mining the cause of political decision-making from social media: A case study of COVID-19 policies across the US states. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 288–301, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021c. Causal direction of data collection matters: Implications of causal and anticausal learning for NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *ACL*.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 656–666.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? A causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

J. Pearl and E. Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In *Proc. AAAI 2011*, pages 247–254.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. 2012. On causal and anticausal learning. In *ICML-12*, Edinburgh, Scotland.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Towards causal representation learning. *CoRR*, abs/2102.11107.

Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors. 2022. *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*. PMLR.

S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A.J. Kerminen. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030.

P. Spirtes, C. Glymour, and R. Scheines. 2001. *Causation, Prediction, and Search*, 2nd edition. MIT Press, Cambridge, MA.

Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. SpringerOpen.

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. In *International Joint Conference on Artificial Intelligence*.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022. A causal framework to quantify the model robustness on math word problems.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*.

Victor Veitch, Dhanya Sridhar, and David M Blei. 2020. Adapting text embeddings for causal inference. In *UAI*.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *EMNLP*.

F. Xie, R. Cai, B. Huang, C. Glymour, Z. Hao, and K. Zhang. 2020. Generalized independent noise condition for estimating linear non-gaussian latent variable causal graphs. In *Neural Information Processing Systems (NeurIPS)*.

Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. 2022. Identification of linear non-Gaussian latent hierarchical structure. In *Proceedings of the 39th International Conference on Machine Learning*, pages 24370–24387.

Weiran Yao, Guangyi Chen, and Kun Zhang. 2022. Causal disentanglement for time series. In *Neural Information Processing Systems (NeurIPS)*.

K. Zhang and A. Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. 2013. Domain adaptation under target and conditional shift. In *ICML-13*.

Kun Zhang, Mingming Gong, Petar Stojanov, Biwei Huang, Qingsong Liu, and Clark Glymour. 2020. Domain adaptation as a problem of inference on graphical models. In *Neural Information Processing Systems (NeurIPS)*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.