Tutorial on Causal Inference for Natural Language Processing

Zhijing Jin
Max Planck Institute & ETH Zürich
zjin@tue.mpg.de

Amir Feder
Technion - Israel Institute of Technology
feder@campus.technion.ac.il

Kun Zhang
Carnegie Mellon University
kunz1@cmu.edu

Abstract

Causal inference is becoming an increasingly important topic in deep learning, with the potential to deal with critical deep learning problems such as representation learning, 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 recent advances in causal inference and causal discovery to the NLP audience, provide an overview of the state-of-the-art causal treatments of NLP problems, and aim to further inspire novel approaches to NLP. We will cover a general introduction to basics of causal discovery and inference, how to use causality to improve NLP models, and how to use causality to improve discovery of linguistic theories and NLP for social applications. This tutorial is inclusive to a variety of audience and is expected to facilitate the community's developments in formulating and addressing new, important NLP problems in light of the emerging causal principles and methodologies.

1 Introduction

Establishing causal relationships is a fundamental goal of scientific research (Pearl, 2000; Spirtes et al., 2001). Quantifying the effectiveness of a vaccine, the persuasive power of a public health ad, or the impact of a lockdown policy, naturally boils down to questions of causality: How would the treatment (vaccine, ad or policy, etc.) affect the outcome (e.g., infection rates) compared to a counterfactual world with no treatment? The direction and strength of causal relationships, once formally identified, play a key role in the formulation of clinical treatments, public policy, and other long-standing prescriptive strategies.

Text plays an increasingly important role in the study of causal relationships across domains. In the setting of crowd lending, communication researchers have examined the impact of the language of funding requests on receiving loans (Larrimore et al., 2011), where text is the treatment. In the context of immigration policy, political scientists have studied the impact of knowing an illegal immigrant's criminal history on written justifications of whether they should be jailed (Egami et al., 2018), where text is the outcome. In the computer science literature, a growing body of literature focuses on the interplay between natural language processing and causal inference (Tan et al., 2014; Wood-Doughty et al., 2018; Sridhar and Getoor, 2019; Veitch et al., 2020; Keith et al., 2020; Zhang et al., 2020a; Feder et al., 2020).

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 are underexplored.

This tutorial will review research at the intersection of causal inference and NLP across research areas, and situate this in the broader NLP landscape. We will discuss research that uses causal inference to improve the learning performance, robustness, fairness, and interpretability of NLP models. We will also explore the full spectrum of causal estimation settings with text as a treatment, control, and outcome (among others). Finally, we aim at providing a unified overview of causal inference for the computational linguistics community.

2 Tutorial Overview

This introductory tutorial aims to introduce causal inference 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 in this tutorial will first focus on providing a generalized introduction to causal inference 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: estimating causal effects from text, and using causal formalisms to make NLP methods more interpretable, robust and fair. Accordingly, we divide the tutorial's content into the following three parts:

1. Introduction to Causality. We will give a broad coverage of central concepts, principles and technical developments in causal modeling, identification of causal effects (sometimes known as causal inference), and how to find 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, recent machine learning methods for causal discovery (Spirtes et al., 2001; Peters et al., 2017; Spirtes and Zhang, 2016; Shimizu et al., 2006; Zhang and Hyvärinen, 2009), and 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, 2020b).

Specifically, we will answer the following questions in this part. 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? What role does causality play in machine learning under data heterogeneity? How can unsupervised deep learning benefit from a causal view? How are robustness and fairness in trustworthy AI connected to causal principles? Is causality a component in AI at a higher-level? If it is, how?

2. Causality for Estimation. We then outline the possibilities and challenges of using NLP to identify causal effects in each case – text as outcome, text as treatment, and text as confounder, and how assumptions and estimation are complicated by the addition of text data. We will demonstrate how to estimate causal effects in the presence of text, in each of these settings.

Many scientific fields are increasingly interested in incorporating text as data (e.g., Roberts et al., 2014; Pryzant et al., 2017; Zhang et al., 2020a). A key property of these fields that may be unfamiliar to NLP researchers is the emphasis on causal inference, often to evaluate policy interventions. For example, before recommending a new drug therapy, clinicians want to know the causal effect of the drug on disease progression. Causal inference involves a question about a counterfactual world created by taking an intervention: what would a patient's disease progression have been if we had given them the drug? As we will explain in this tutorial, in observational data, the causal effect is not equivalent to the correlation between patients taking the drug and their observed disease progression. There is now a deep literature on techniques for making valid inferences using traditional (non-text) datasets (e.g., Morgan and Winship, 2015), but the application of these techniques to natural language data raises new and fundamental challenges.

3. Causality for Prediction. Finally, we consider the converse relationship: using causal reasoning to help solve more traditional NLP tasks like understanding, manipulating, and generating natural language. Despite its critical role in the life and social sciences, causality has not had the same importance in NLP. At first glance, NLP may appear to have little need for causal ideas. The remarkable progress that the field has achieved in the past few years has been derived from the use of increasingly high-capacity neural architectures to extract correlations from large-scale datasets (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019). Yet despite the march of state-of-the-art results, correlational predictive models can be untrustworthy (Jacovi et al., 2021): they may latch onto spurious correlations, leading to errors in out-of-distribution settings (e.g., McCoy et al., 2019); they may exhibit unacceptable performance differences across groups of users (e.g., Zhao et al., 2017); and their behavior may be too inscrutable to incorporate into high-stakes decisions (Guidotti et al., 2018).

In this part, we will demonstrate how each of these shortcomings can potentially be addressed by the causal perspective: knowledge of the causal re-

lationship between observations and labels can be used to formalize spurious correlations and to mitigate predictor reliance on them (Bühlmann, 2020; Veitch et al., 2021); causality also provides a language for specifying and reasoning about fairness conditions (Kilbertus et al., 2017); the task of explaining predictions may be naturally formulated in terms of counterfactuals (Feder et al., 2021b).

3 Tutorial Outline

For the three-hour tutorial, we will cover three main topics: introduction of causality, how causality can help improve NLP models, and how causality can help linguistic theories and NLP for social applications. Each of these three topics will have a 45-min lecture and a 5-min break following the lecture. The remaining 30 min will be used for an interactive exercise with the audience and Q&A.

An outline of the tutorial content is as follows:

1. Introduction to causality (45-min lecture + 5-min break)
    - Motivations: Why is causality helpful for NLP? What subtopics are the most applicable to NLP?
    - Basics of causal discovery and causal inference
2. Causality to help improve NLP models (45-min lecture + 5-min break)
    - Coverage of a list of areas where causality can help improve NLP models, including model robustness, domain adaptation, debiasing models, interpretability, and fairness.
3. Causality to help linguistic theories and NLP for social applications (45-min lecture + 5-min break)
    - Cases where causal discovery and inference can help verify linguistic theories
    - 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.
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
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 Schoelkopf, Clark Glymour, and Peter Spirtes. Apart from these causality researchers' work, 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 (Stanford), and many others.

5 Organizing Committee

Zhijing Jin (she/her) is a PhD at Max Planck Institute and ETH Zürich supervised by Prof Bernhard Schoelkopf. Her research aims to (1) improve NLP models by connecting NLP with causal inference, and (2) expand the impact of NLP by promoting NLP for social good. 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 is the Publications Chair for the 1st conference on Causal Learning and Reasoning (CLeaR). She is also actively involved in AI for social good, as the organizer of NLP for Positive Impact Workshop at ACL 2021, and RobustML workshop at ICLR 2021. To support the NLP research community, she organizes the ACL Year-Round Mentorship Program.

Amir Feder (he/him) is a PhD at the Technion - Israel Institute of Technology, working with Prof Roi Reichart and Prof Uri Shalit. 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 the organizer of the First Workshop on Causal Inference and NLP (CI+NLP) at EMNLP 2021.

Kun Zhang (he/him) is an associate professor at Carnegie Mellon University. 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 Learning, 2020, co-editor of a number of journal special issues on causality, and many others.

6 Diversity Efforts

We have taken various steps to ensure that our tutorial is inclusive and diverse, and will continue these efforts if the workshop is accepted.

Our organizing committee include 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. Members of these committees have also been actively involved in promoting diversity and inclusivity in the NLP community.

The topic of our workshop is foundamentals of causal inference, which is a helpful tool across many NLP tasks, and the methods can scale up to various languages and domains.

If accepted, we will advertise the tutorial to diversity-oriented venues (e.g., Widening NLP, QueerInAI, BlackInAI, WiML). To encourage participants from diverse research backgrounds, we will also advertise in non-NLP areas related to NLP and causal inference, and seek out funding for travel grants for underrepresented minorities. We will also have an archival and non-archival track, to encourage work from those fields with presentation restrictions.

7 Target Audience & Prerequisites

There is no required audience background. Preferred knowledge include 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 set up of simple NLP tasks such as classification).

Estimated Audience Size: This is the first tutorial on NLP and causal inference. As a reference, a similar workshop, the first workshop on Causal Inference + NLP (CI+NLP), will be held at EMNLP 2021. Given the increasing interest in causal inference techniques for NLP, we expect 50-200 attendees to this tutorial.

8 Virtual Organizing & Technical Needs

We are ready to host this workshop in either a hybrid or a virtual setting (if in-person isn't possible), and we will draw from our previous experience organizing virtual tutorials, and presenting and attending virtual conferences, to ensure that the virtual workshop runs smoothly.

We will not require any special features other than standard live conferencing (virtual) or a room with internet and a projector (in-person).

9 Recommended Reading List

We compiled a recommended reading list of causality and NLP papers at `https://github.com/zhijing-jin/Causality4NLP_Papers`
Among the papers, the top three recommended readings are Guo et al. (2020), Schölkopf et al. (2021) and Feder et al. (2021a).

10 Other Information

Tutorial Type: Introductory.
Tutorial Materials: We will make available all tutorial presentation materials, including slides, captioned video recording, codes, and the recommended paper list.

11 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.

References

Peter Bühlmann. 2020. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

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. 2021a. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *CoRR*, abs/2109.00725.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2020. CausaLM: Causal model explanation through counterfactual language models. *arXiv preprint arXiv:2005.13407*.

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

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.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635.

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.

Laura Larrimore, Li Jiang, Jeff Larrimore, David Markowitz, and Scott Gorski. 2011. Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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*.

Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Reid Pryzant, Youngjoo Chung, and Dan Jurafsky. 2017. Predicting sales from the language of product descriptions. In *eCOM@ SIGIR*.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

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.

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*.

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*.

Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020a. Quantifying the causal effects of conversational tendencies. In *CSCW*.

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. 2020b. 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.