Identifying Causal Influences on Publication Trends and Behavior: A Case Study of the Computational Linguistics Community

Maria Glenski and **Svitlana Volkova**

National Security Directorate

Pacific Northwest National Laboratory

{Maria.Glenski@pnnl.gov, Svitlana.Volkova@pnnl.gov}

Abstract

Drawing causal conclusions from observational real-world data is a very much desired but challenging task. In this paper we present mixed-method analyses to investigate causal influences of publication trends and behavior on the adoption, persistence, and retirement of certain research foci – methodologies, materials, and tasks that are of interest to the computational linguistics (CL) community. Our key findings highlight evidence of the transition to rapidly emerging methodologies in the research community (*e.g.,* adoption of bidirectional LSTMs influencing the retirement of LSTMs), the persistent engagement with trending tasks and techniques (*e.g.,* deep learning, embeddings, generative, and language models), the effect of scientist location from outside the US, *e.g.,* China on propensity of researching languages beyond English, and the potential impact of funding for large-scale research programs. We anticipate this work to provide useful insights about publication trends and behavior and raise the awareness about the potential for causal inference in the computational linguistics and a broader scientific community.

Causal understanding is essential for informed decision making (Pearl and Mackenzie, 2018; Pearl, 2019; Varian, 2016) to go beyond correlations and overcome the predictability limit of real-world partially observed systems including complex systems of human social behavior (Abeliuk et al., 2020; Hofman et al., 2017).

Unlike earlier work that focused on analysing publication trends, diversity and innovation in science relying on descriptive exploratory analysis primarily driven by correlations (Fortunato et al., 2018; Hofstra et al., 2020; Ramage et al., 2020), this work aims to provide empirical evidence of causal mechanisms driving publication trends and behavior in the computational linguistics community. Our key contributions are two-fold. First,

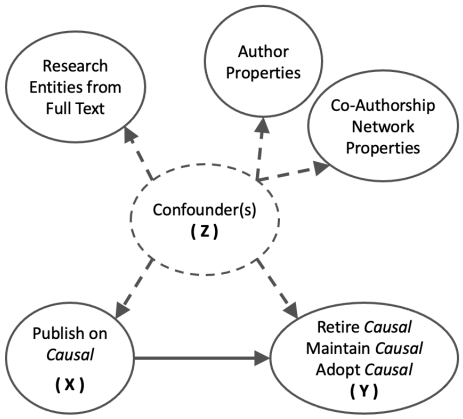

Figure 1: Causal diagram for scientific publications, where full text, author, and co-citation network properties encode causal confounders (e.g., author influence, research tasks and methodology) used as covariates. Our analyses rely on the assumption that latent confounders can be measured and adjusted for based on proxies in our covariates.

we experiment and evaluate the potential of using complementary causal inference approaches, specifically causal structure learning models and several treatment effect estimation techniques to measure causal influences in high-dimensional observational data. Second, we analyze the temporal dynamics of causal influences of publication trends on the adoption, persistence and retirement of certain research foci in the CL community.

For our analyses we leverage public ACL anthology data and encode publication and scientist characteristics, as well as collaboration behavior as confounders, to measure the effect of previous research foci on the adoption, maintenance and retirement of future research foci focusing on most recent six years between 2014 and 2019. Figure 1 presents a causal diagram which illustrates our core assumption: that latent confounders (*e.g.,* author reputation, productivity, and collaborations, the strength and length of authors' research careers, novelty of research, papers' contributions to the

field, etc.) can be measured and adjusted for based on proxies in our covariates (*e.g.,* research entities extracted from publications, author properties including time since first paper, volume of papers, centrality within co-authorship networks etc.).

1 Related Work

There are two complementary causal inference frameworks – structural causal models (Pearl et al., 2009) and treatment effect estimation approaches (Rosenbaum and Rubin, 1983). Existing approaches to learn the causal structure (aka causal discovery) broadly fall into two categories: constraint-based (Spirtes et al., 2000; Yu et al., 2016) and score-based (Chickering, 2002).

Recently, there have been an increased interest in causal inference on observational data (Guo et al., 2020), including text data, in the computational linguistics and computational social science communities (Lazer et al., 2009). For example, recent work by (Roberts et al., 2020) estimated the effect of perceived author gender on the number of citations of the scientific articles and (Veitch et al., 2020) measured the effect that presence of a theorem in a paper had on the rate of the paper's acceptance.

Additional examples in the computational social science domain include: measuring the effect of alcohol mentions on Twitter on college success (Kiciman et al., 2018); estimating the effect of the "positivity" of the product reviews and recommendations on sales on Amazon (Pryzant et al., 2020; Sharma et al., 2015); understanding factors effecting user performance on StackExchange, Khan Academy, and Duolingo (Alipourfard et al., 2018; Fennell et al., 2019); estimating the effect of censorship on subsequent censorship and posting rate on Weibo (Roberts et al., 2020) and word use in the mental health community on users' transition to post in the suicide community on Reddit (De Choudhury et al., 2016; De Choudhury and Kiciman, 2017); or the effect of exercise on shifts of topical interest on Twitter (Falavarjani et al., 2017). Moreover, Keith et al. (2020) presented the overview of causal approaches for computational social science problems focusing on the use of text to remove confounders from causal estimates, which was also used by (Weld et al., 2020). Earlier work utilized matching methods to learn causal association between word features and class labels in document classification (Paul, 2017; Wood-Doughty et al., 2018), and use text as treatments (Fong and Grimmer, 2016; Dorie et al., 2019) or covariates e.g., causal embeddings (Veitch et al., 2020; Schölkopf et al., 2021).

Several studies have leveraged the ACL Anthology dataset to analyze diversity in computer science research (Vogel and Jurafsky, 2012), and perform exploratory data analysis such as knowledge extraction and mining (Singh et al., 2018b; Radev and Abu-Jbara, 2012; Gábor et al., 2016). However, unlike any other work, our approach focuses on leveraging complementary methods for causal inference – structural causal models and treatment effect estimation to discover and measure the effect of scientists' research focus on their productivity and publication behavior, specifically the emergence, retirement and persistence of computational linguistics methodologies, approaches and topics.

2 Data Preprocessing

Our causal analysis relies on the publication records from the Association of Computational Linguistic (ACL) research community from 1986 through 2020. We collect the ACL Anthology dataset[1] (Gildea et al., 2018) with the bibtex provided with the accompanying abstracts. Excluding all records that do not contain authors (*e.g.,* bibtex entries for the workshop proceedings), we convert the bibtex representation into a data representation where each row represents each paper-author combination (*i.e.,* for a paper (paperX) with three authors, there are three representative rows: paperX-author1, paperX-author2, and paperX-author3).

Then, we extract features that encode **paper** properties: the year it was published, whether the paper was published in a conference or journal, the number of authors, the number of pages, and word count in paper. We also compute Gunning fog index (Gunning et al., 1952) – influenced by the number of words, sentences, and complex words.

We then annotate each row with properties related to the **author** during the year the paper was published. As a proxy of the length of the author's research career in the computational linguistics community, we calculate the number of years since the author's first publication in the anthology. Each author's location is represented as the location (country) of the institution the author is associated with in the metadata or full text. To measure productivity at varying granularities, we calculate the number of one's papers published in

[1] https://aclanthology.org/

total, in the last year, and in the last five years.

We then construct a dynamic network representation of the anthology using author-to-paper relationships for each calendar year, as encoded in the metadata. After projecting those relationships into the dynamic co-authorship network that reflects author to co-author connections by year, we calculate centrality and page rank network statistics over time to measure the influence of the author. These **collaboration behavior** features complement previously described author properties. We also added three features to encode the diversity in co-authorship. First, the number of all co-authors who published the papers with the author. Second, the average number of papers co-authored per co-author, which is computed as the total number of papers co-authored per co-author divided by the number of co-authors. The last is a likelihood that a co-author is an author on a paper, which is the second feature divided by the total number of the author's papers. This enables us to measure the diversity, or lack thereof, of collaborative relationships of each author, and encodes how collaboration behavior evolves over time.

2.1 Encoding Research Focus

After extracting the full text of each paper from the PDF using GROBID (GRO, 2008–2021), we use the SpERT model trained to extract key research entities from scientific publications. The SpERT model (Luan et al., 2018) was trained to extract scientific entities of different types such as tasks, methods, and materials and the relationships between them such as "Feature-Of" and "Used-for", using the SciERC dataset[2]. After applying the model to the ACL data, we consolidate noisy references of research entities into representative clusters manually, resulting in 50 entities that encode research tasks, methods, and materials[3].

[2] http://nlp.cs.washington.edu/sciIE/

[3] Research entities trending in the CL community used for our causal analyses: "artificial intelligence", "adversarial", "annotation", "arabic", "attention", "baselines", "bidirectional lstm", "causal", "chinese", "classification", "coreference", "crowdsourcing", "deep learning", "dialog", "embeddings", "ethics", "explanation", "fairness", "french", "generative", "german", "grammars", "graph models", "heuristics", "interpretability", "language models", "lstm", "machine learning", "monolingual", "multilingual", "multiple languages", "NER", "node2vec", "non-English language", "pos/dependency/parsing", "QA", "reinforcement learning", "robustness", "russian", "sentiment", "statistical/probabilistic models", "summarization", "topic model", "transfer learning", "transformers", "translation", "transparency", "unsupervised methods", "word2vec", "benchmark".

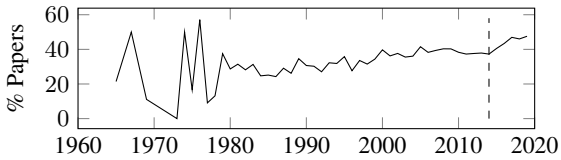

Figure 2: Relative coverage of consolidated research entity representations in the ACL data. Percentage of papers with at least one entity associated by publication year. Dashed line indicates the start of our causal analysis period (2014).

These consolidated entities are representative of the top 300 entities extracted from all ACL anthology publications for which we were able to extract the full text (121,134 out of 127,041 which is 95.3% of all records for which there was an ACL anthology bibtex entry), after removing trivial or general terms such as "system", 'approach", "it", "task", and "method". We present the coverage across papers (% of papers with at least one associated entity) over time in Figure 2, illustrating the coverage approximates the overall coverage (around 41%) for the bulk of the dataset (1980-2019), with coverage trending upwards over time.

3 Methodology

We investigate the causal relationships between characteristics of individual publications for each researcher who authored at least two publications hosted in the ACL Anthology. We identify causal influences of publication outcomes related to the adoption, retirement, and maintenance of computational linguistic methodologies, approaches, or topics, as well as the productivity of authors using both causal discovery and pairwise causal effect estimation.

3.1 Treatments and Outcomes

We consider the inclusion of our consolidated research entities to be treatments in our analyses — *e.g.,* does the inclusion of *biLSTM* architectures in the publication have a causal relationship with future research outcomes? — and the basis of several research outcomes related to the *adoption*, *retirement*, and *maintenance* of CL methodologies, tasks and approaches.

That is, the association of the identified research entities with authors' publications allows us to identify when authors adopt new emerging technologies (*e.g.,* the first use of *transformers*), retire previously used methods or research applications (*e.g.,* if au-

thors stop publishing on *LSTM* architectures after *biLSTM* architectures are introduced), continue to use – or maintain publications in – methods (*e.g.,* when authors continue to publish on *NER*). We associate these behaviors as future outcomes for each author's publications in previous years.

For each year in which an author published in an ACL venue, we calculate adoption and retirement outcomes for each consolidated research element the following year, maintenance outcomes for each research element considering the following two years. Alongside these fine-grained research outcomes, we also examine coarse-grained, or general outcomes for authors:

- overall pauses in publishing within ACL venues (no publications in any ACL community for two years),
- persistent publication records (continuing to publish in consecutive years),
- publication volume increases (the increase or decrease in number of publications in ACL venues relative to the previous year).

In our analyses, we focus on recent six years (2014-2019) for which we have complete treatment and outcome annotations and consider each year independently. We leverage two types of publication record granularities – publication records and yearly research portfolios – to analyze the temporal dynamics of the causal system underpinning CL publication venues at multiple resolutions. Note, we present a detailed description of the treatments, covariates and outcomes we used in Appendix A.

3.2 Causal Structure Learning

Structural causal models are a way of describing relevant features of the world and how they interact with each other. Essentially, causal models represent the mechanisms by which data is generated. The causal model formally consists of two sets of variables U (exogenous variables that are external to the model) and V (endogenous variables that are descendants of exogenous variables), and a set of functions f that assign each variable in V a value based on the values of the other variables in the model. To expand this definition: a variable X is a direct cause of a variable Y if X appears in the function that assigns Y value. Graphical models or Directed Acyclic Graphs (DAGs) have been widely used as causal model representations.

The causal effect rule is defined as: given a causal graph G in which a set of variables $PA(X)$ are designated as a parents of X, the causal effect of X on Y is given by:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, PA = z)P(PA = z), \quad (1)$$

where z ranges over all the combinations of values that the variables in PA can take.

The first approach for our causal analysis aims to examine the causal relationships that are identified using an ensemble of causal discovery algorithms (Saldanha et al., 2020). Our ensemble considers the relationships identified by CCDR (Aragam and Zhou, 2015), MMPC (Max-Min Parents-Children) (Tsamardinos et al., 2003), GES (Greedy Equivalence Search) (Chickering, 2002), and PC (Peter-Clark) (Colombo and Maathuis, 2014). We use the implementations provided by the pcalg R package (Hauser and Bühlmann, 2012; Kalisch et al., 2012) and causal discovery toolbox (CDT) (Kalainathan and Goudet, 2019)[4]. The outcomes of our ensemble approach to causal discovery is a causal graph reflecting the relationships within the causal system, weighting edges by the agreement among the individual algorithms on whether the causal relationship exists.

After applying this causal discovery approach to each year individually, we are able to construct a dynamic causal graph and investigate trends in causal relationships – *e.g.,* as they are introduced, persist over time, or are eliminated.

3.3 Treatment Effect Estimation

We further investigate the magnitude and effect of causal relationships using average treatment effect (ATE) estimates. We compare pair-wise estimates using several causal inference models: *Causal Forest* (Tibshirani et al., 2018) and *Propensity Score Matching* (Ho et al., 2007) using the "MatchIt" R package[5], and a cluster-based conditional treatment effect estimation tool – *Visualization and Artificial Intelligence for Natural Experiments (VAINE)*[6] (Guo et al., 2021).

VAINE is designed to discover natural experiments and estimate causal effects using observational data and address challenges traditional approaches have with continuous treatments and high-

[4] https://fentechsolutions.github.io/CausalDiscoveryToolbox/html/index.html

[5] https://cran.r-project.org/web/packages/MatchIt/vignettes/MatchIt.html

[6] https://github.com/pnnl/vaine-widget

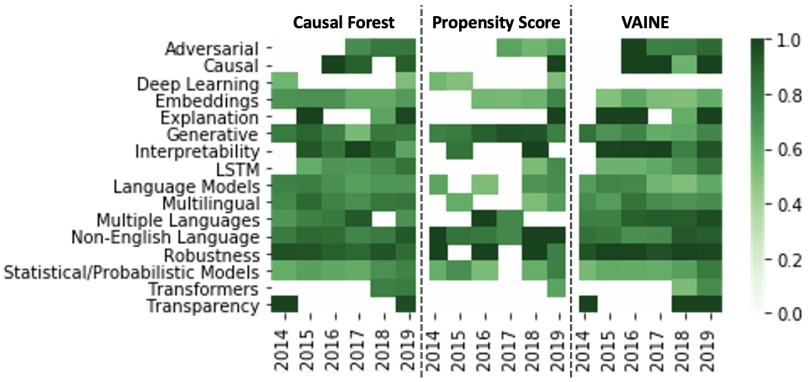

Figure 3: Treatment effect estimates obtained using three causal inference methods – Causal Forest, Propensity Score Matching and VAINE, for publish on $x \rightarrow$ retire x over time, across TEE methods.

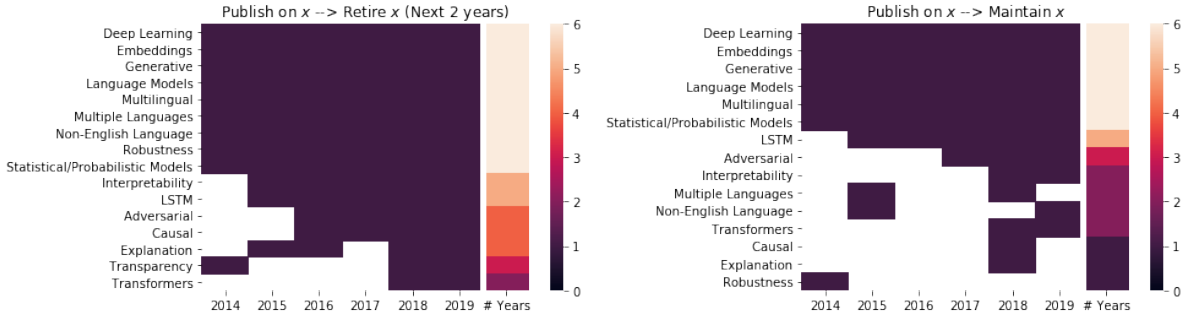

Figure 4: Summary of causal structure learning using our ensemble model discovered from Publish on x to Retire x (above) or Maintain x in the next 2 years (below), by year. Shaded cells indicate that an edge was discovered, white cells indicate that no edge was discovered for that year. At right is a summary of the number of years for which an edge was discovered.

dimensional feature spaces. First, VAINE allows users to automatically detect sets of observations controlling for various covariates in the latent space. Then, using linear modeling, VAINE allows to estimate the treatment effect within each group and then average these local treatment effects to estimate the overall average effect between a given treatment and outcome variable. VAINE's novel approach for causal effect estimation allows it to handle continuous treatment variables without arbitrary discretization and produces results that are intuitive, interpretable, and verifiable by a human. VAINE is an interactive capability that allows the user to explore different parameter settings such as the number of groups, the alpha threshold to identify significant effects, etc.

Below we define what we mean by learning a causal effect from observational data. Given n instances $[(x_1, t_1), \ldots, (x_n, t_n)]$ learning causal effects quantifies how the outcome y is expected to change if we modify the treatment from c to t, which can be defined as $\mathbb{E}(y \mid t) - \mathbb{E}(y \mid c)$, where t and c denote a treatment as a control.

Similarly to our causal discovery based analyses, we examine the growth and decay of causal influence for a series of treatments (research focus represented by materials, methodology, or application-based keywords) on our outcomes of interest.

3.4 Evaluation

Evaluating causal analysis methods is challenging (Saldanha et al., 2020; Weld et al., 2020; Gentzel et al., 2019; Shimoni et al., 2018; Mooij et al., 2016; Dorie et al., 2019; Singh et al., 2018a; Raghu et al., 2018). Broadly, evaluation techniques include structural, observational, interventional and qualitative techniques e.g., visual inspection. Observational evaluation by nature are non-causal and do not have the ability to measure the errors under interventions. Structural measures are limited due to the requirements of known structure, are oblivious to magnitude and type and dependence, as well as treatments and outcomes, and constrain research directions. Unlike structural and observational measures, interventional measures allow to evaluate model estimates of interventional effects e.g., "what-if counterfactual evaluation".

In this work we rely on both qualitative and quantitative evaluation. Methods that we use for causal inference were independently validated using structural and observational measures on synthetic datasets – causal forest (Wager and Athey, 2018), propensity score matching (Cottam et al., 2021), causal ensemble (Saldanha et al., 2020), and VAINE (Guo et al., 2021). Since we rely on four complementary causal inference techniques, we draw our conclusions based on their agreement. In addition, to perform qualitative evaluation with the human in the loop we rely on recently release visual analytics tools to evaluate causal discovery and inference (Cottam et al., 2021).

4 Results

In this section, we present a series of key findings surfaced within the causal mechanisms discovered and treatment effects estimated focusing on contrastive analysis over time: whether publishing on a given research entity (methodology, task, material) influences continuing to publish in that area or with that methodology, how authors shift from existing to novel methodology over time, and evidence of external events (*i.e.,* funding or large research programs) potentially impacts the adoption and maintenance of publication trends.

4.1 Continuing Existing Avenues of Research

One of the first trends we noticed, in both the causal structures and treatment effects, was a causal relationship between publishing on a given research entity (*e.g.,* robustness, LSTMs, transformers, NER, etc.) and whether an author would *continue* to publish on the same topic, task, or methodology in the following year(s). Does publishing once influence whether you will publish again? In short, no. We see a consistent trend in *positive* treatment effects, as illustrated in Figure 3, from publishing in the current year to not publishing (pausing or retiring research entities) in the future – publishing on x leads to *not publishing* on x in the future.

In Figure 4, we summarize the temporal dynamics of causal relationships from publishing on x in a current year's publication to retiring x (no publications associated with research entity x) in the next 2 years (above) or maintaining x (at least one publication associated) in the next 2 years (below) indicated by our causal structural learning analyses which aligns with our TEE results. We show the consistency in which research entities are included

Method	2014	2015	2016	2017	2018	2019
CF	0	0.71	-0.01	0.07	0.09	-0.03
VAINE	0	0.88	0.46	0.68	0.68	0.77
Mean	0	0.80	0.22	0.32	0.39	0.37

Table 1: Treatment effect estimates for the treatment *Publish on bidirectional LSTM* on outcome *Retire LSTM* by year, illustrating a decaying influence.

Method	2014	2015	2016	2017	2018	2019
CF	0	0.76	-0.03	0.36	-0.2	0.00
VAINE	0	0	0	0.39	-0.23	0
Mean	0	0.38	-0.02	0.38	-0.22	0.00

Table 2: Treatment effect estimates for the treatment *Publish on bidirectional LSTM* on outcome *Increase Publications next year*, illustrating a strong initial influence shift to negative (2018) then neutral (2019).

in these trends using Figure 5. We see that many of the elements where causal relationships were identified in all 6 years are present in both the retirement and maintenance relationships. Of all the elements, research on *Transparency* is the only case where there is only a retirement relationship. All elements with identified maintenance relationships in at least one year were also present in the set of retirement relationships.

4.2 Emerging Research Foci, and the Impacts on Retirement of Old Research Foci

The introduction or popularization of new model architectures (especially in deep learning) has an initial strong impact on retirement of previous SOTA architectures, but this is often focused on the initial adoption. We investigate several examples of such phenomena. Table 2 illustrates the decaying causal influence that using bidirectional LSTM-based architectures in current publications has on the retirement of (no longer using) LSTM in future publications. At first, there is a strong causal effect (approx. 0.8), where the use of biLSTM layers lead to no longer using LSTM layers. However, this reduces over time, with CF estimating close to no effect past 2015. We see a complementary trend on the relative publication volume increase outcome (*Increase # publications next year*), where there is an initial strong effect (0.76) that decays until it shifts to a negative effect (in 2018) then neutral (in 2019), as shown in Table 2.

In addition, we see a consistent divergence from the trend described above (publishing on x influ-

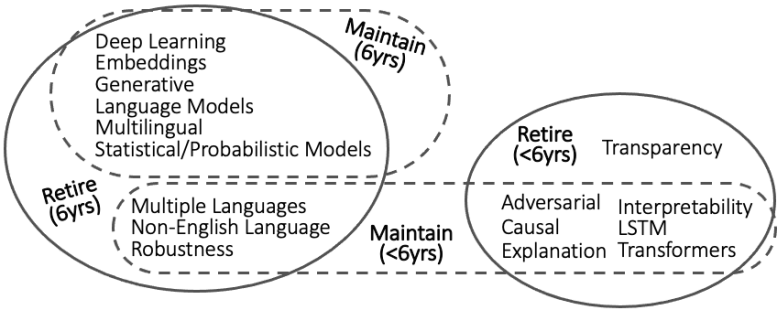

Figure 5: Venn diagram illustrating shared research focus among causal relationships discovered by multiple causal inference methods in Figure 4. This plot demonstrates long-lasting vs. short-lasting research trends in the CL community.

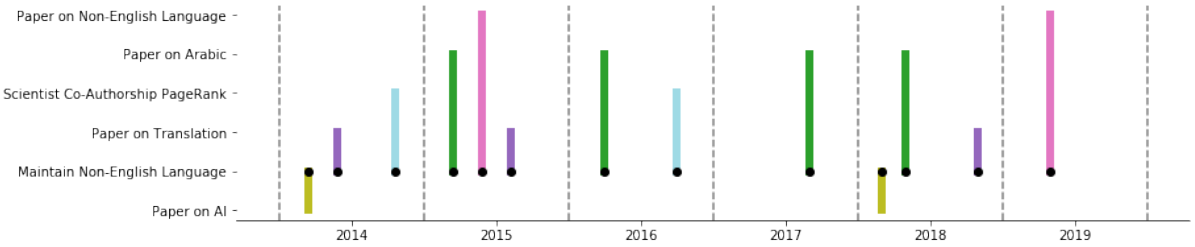

Figure 6: Recurrent causal relationships (identified for at least two years) that influence **continued publication patterns** related to non-English languages in the CL community, *e.g.,* scientist co-authorship PageRank effects maintaining non-English publication focus in 2014 and 2016. Black markers identify the effect, with line segments extending to the cause nodes, and distinct relationships are represented by varying colors.

ences not publishing on x in the next two years) for research related to non-English languages in the 2016-2018 time frame (see Figure 6). These might be explained by the impact of large-scale research programs and funding (note, we will empirically confirm or dispute this hypothesis as a part of our future work). For example, we find that these outcomes (whether researchers continue to publish research related to non-English languages in 2017-2020) align with the last few years (and program-wide evaluation events[7]) of the LORELEI (Low Resource Languages for Emergent Incidents) DARPA program[8].

The goal of the LORELEI program was "to dramatically advance the state of computational linguistics and human language technology to enable rapid, low-cost development of capabilities for low-resource languages", and resulted in several publications on such languages from performers *e.g.,* (Strassel and Tracey, 2016). "What is funded is published" may be an intuitive influence, but here we see qualitative evidence that funding could influence the causal mechanisms of the publication ecosystem — these signals are strong enough to be reflected in causal systems discovered using causal discovery algorithms in observational data. For *adopting* non-English as a research focus, we also see influence from the authors' country associations – *i.e.,* institution affiliations in China influence adopting non-English research (Fig. 7).

Table 3 illustrates a peak in the positive influence of publishing in a particular non-English language (*i.e.,* Arabic, which was one of the languages of interest for the LORELEI program) and continuing to publish on non-English languages. We see that the divergence of the causal relationships, and the persistence of authorship in non-English language research, illustrated in Figure 8, center around or peak in 2017 and begin to flip (causal forest estimates a negative effect of -0.55) in 2019.

5 Discussion and Conclusions

In this study, we identified and analyzed causal influences between the trends and publishing behavior on future research performed and published in the computational linguistic community. Adjusting for confounders related to publication properties extracted from publication text and metadata, author characteristics, and collaboration network proper-

[7] https://www.nist.gov/itl/iad/mig/lorehlt-evaluations

[8] https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

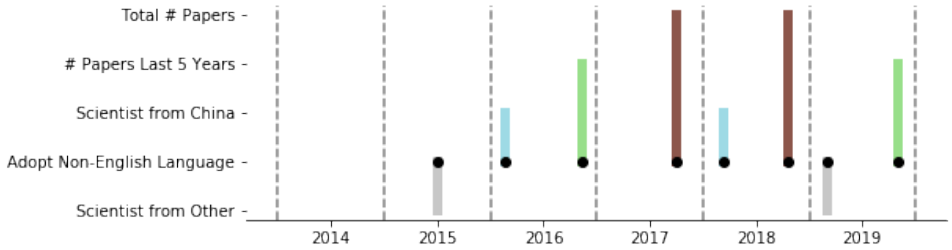

Figure 7: Recurrent causal relationships that influence **new publication patterns** related to non-English languages.Black markers identify the effect, with line segmnets extending to the cause nodes, and distinct relationships are represented by varying colors. Empty time windows indicate no recurrent relationships were discovered.

Publish on Arabic → Continue Publishing on non-English

Method	2014	2015	2016	2017	2018	2019
CF	0.04	0.03	0.28	0.56	0.40	-0.55
VAINE	0	0.20	0.43	0.51	0.13	0.06
Mean	0.02	0.12	0.36	0.54	0.27	0.00

Publish on Arabic → Stop Publishing on non-English

Method	2014	2015	2016	2017	2018	2019
CF	0.44	0.80	0.40	0.13	0.50	0.91
VAINE	0.81	0.71	0.45	0.20	0.82	0.91
Mean	0.62	0.75	0.42	0.16	0.66	0.91

Publish on Arabic → Increase Publications next year

Method	2014	2015	2016	2017	2018	2019
CF	-0.03	-0.03	0.31	1.8	0.23	-0.09

Table 3: Treatment effect estimates for the treatment *Publish on Arabic* for outcomes *Maintain non-English in the next two years* (above) and *Stop publishing on non-English in the next year* (below) by year, illustrating a peak in influence continuing to publish in 2017.

ties, we examine the causal relationships that could potentially be driving scientist productivity and the adoption, maintenance, and retirement of methods, materials, and tasks referenced in publications, and how these dynamics evolve over time.

Our analyses show that publishing once on a specific task, application, or methodology has a causal influence that causes authors not to publish on the same approach in the following year e.g., robustness, interpretability etc. This is consistent across a significant number of the research methods, tasks, and application domains represented in our consolidated annotations.

There are several potential drivers of this causal relationship. First, publishing in the CL community, particularly in recent years, is extremely competitive. Acceptance rates are low and the number of submissions each year continue to grow – there is a lot of competition for few spots. This

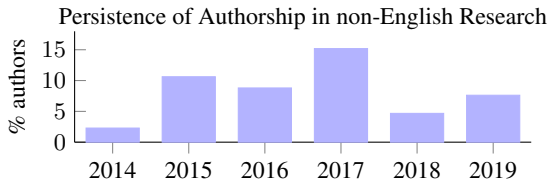

Figure 8: The percentage of authors who published on non-English languages in a given year, who also published on non-English languages in the following year.

could also be reflective of the churn in novelty and state-of-the-art technologies: as one technology (*e.g.,* LSTMs) are replaced by a new SOTA methodology (*e.g.,* biLSTMs, or transformers), this naturally leads to the retirement of the previous methods, which we also see reflected in the findings presented in subsection 4.2. As we drive the field forward, we stop publishing on older methods, materials, and tasks because of novelty incentives or requirements to be accepted in top-tier venues.

Limitations A limitation of our current causal analysis approach is the restriction to ACL Anthology publication records only. As this dataset comprises only ACL venues, it does not guarantee inclusion of all possible publications from each author. For example, authors who publish in ACL venues may also publish in ICLR, NeurIPs, etc. Future work can address this limitation by augmenting the data set with supplementary collects from these venues (*e.g.,* using Google Scholar).

Another avenue of future work is to incorporate funding information directly in the analyses. As we have shown, there is evidence that funding has causal relationships with publication outcomes and expertise evolution, and may act as a confounder for other relationships. Future work will extract funding from full text PDFs (*e.g.,* from the acknowledgements section) in order to adjust for the effects of funding as a confounder, which may also be an impactful treatment to analyze.

Acknowledgements

This material is based on work funded by the United States Department of Energy (DOE) National Nuclear Security Administration (NNSA) Office of Defense Nuclear Nonproliferation Research and Development (DNN R&D) Next-Generation AI research portfolio and Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO1830. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Government or any agency thereof. We would like to thank Joonseok Kim and Jasmine Eshun for their assistance preparing data.

References

2008–2021. Grobid. https://github.com/kermitt2/grobid.

Andrés Abeliuk, Zhishen Huang, Emilio Ferrara, and Kristina Lerman. 2020. Predictability limit of partially observed systems. *Scientific reports*, 10(1):1–10.

Nazanin Alipourfard, Peter G Fennell, and Kristina Lerman. 2018. Using simpson's paradox to discover interesting patterns in behavioral data. In *Twelfth International AAAI Conference on Web and Social Media*.

Bryon Aragam and Qing Zhou. 2015. Concave penalized estimation of sparse gaussian bayesian networks. *The Journal of Machine Learning Research*, 16(1):2273–2328.

David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Diego Colombo and Marloes H Maathuis. 2014. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782.

Joseph Cottam, Maria Glenski, Yi Shaw, Ryan Rabello, Austin Golding, Svitlana Volkova, and Dustin Arendt. 2021. Graph comparison for causal discovery. *Visualization in Data Science 2021*.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. 2019. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68.

Seyed Mirlohi Falavarjani, Hawre Hosseini, Zeinab Noorian, and Ebrahim Bagheri. 2017. Estimating the effect of exercising on users' online behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Peter G Fennell, Zhiya Zuo, and Kristina Lerman. 2019. Predicting and explaining behavioral data with structured feature space decomposition. *EPJ Data Science*, 8(1):23.

Christian Fong and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1600–1609.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science*, 359(6379).

Kata Gábor, Haifa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the ACL anthology corpus for the automatic analysis of scientific literature. In *LREC 2016*.

Amanda Gentzel, Dan Garant, and David Jensen. 2019. The case for evaluating causal models using interventional measures and empirical data. *Advances in Neural Information Processing Systems*, 32:11722–11732.

Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. 2018. The ACL anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28.

Robert Gunning et al. 1952. Technique of clear writing.

Grace Guo, Maria Glenski, ZhuanYi Shaw, Emily Salandha, Alex Endert, Svitlana Volkova, and Dustin Arendt. 2021. Vaine: Visualization and ai for natural experiments. *IEEE transactions on visualization and computer graphics*.

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

Alain Hauser and Peter Bühlmann. 2012. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464.

Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236.

Jake M Hofman, Amit Sharma, and Duncan J Watts. 2017. Prediction and explanation in social systems. *Science*, 355(6324):486–488.

Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. 2020. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291.

Diviyan Kalainathan and Olivier Goudet. 2019. Causal discovery toolbox: Uncover causal relationships in python.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. 2012. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344.

Emre Kiciman, Scott Counts, and Melissa Gasser. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *Twelfth International AAAI Conference on Web and Social Media*.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Social science. computational social science. *Science (New York, NY)*, 323(5915):721–723.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204.

Michael Paul. 2017. Feature selection as causal inference: Experiments with text classification. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 163–172.

Judea Pearl. 2019. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books, New York City, NY.

Judea Pearl et al. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2020. Causal effects of linguistic properties. *arXiv preprint arXiv:2010.12919*.

Dragomir Radev and Amjad Abu-Jbara. 2012. Rediscovering ACL discoveries through the lens of acl anthology network citing sentences. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 1–12.

Vineet K Raghu, Allen Poon, and Panayiotis V Benos. 2018. Evaluation of causal structure learning methods on mixed data types. *Proceedings of machine learning research*, 92:48.

Daniel Ramage, Christopher D Manning, and Daniel A McFarland. 2020. Mapping three decades of intellectual change in academia. *arXiv preprint arXiv:2004.01291*.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Emily Saldanha, Robin Cosbey, Ellyn Ayton, Maria Glenski, Joseph Cottam, Karthik Shivaram, Brett Jefferson, Brian Hutchinson, Dustin Arendt, and Svitlana Volkova. 2020. Evaluation of algorithm selection and ensemble methods for causal discovery.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Amit Sharma, Jake M Hofman, and Duncan J Watts. 2015. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 453–470.

Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. 2018. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*.

Karamjit Singh, Garima Gupta, Vartika Tewari, and Gautam Shroff. 2018a. Comparative benchmarking of causal discovery algorithms. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 46–56.

Mayank Singh, Pradeep Dogga, Sohan Patro, Dhiraj Barnwal, Ritam Dutt, Rajarshi Haldar, Pawan Goyal, and Animesh Mukherjee. 2018b. Cl scholar: The acl anthology knowledge graph miner. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 16–20.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.

Stephanie Strassel and Jennifer Tracey. 2016. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280.

Julie Tibshirani, Susan Athey, Rina Friedberg, Vitor Hadad, David Hirshberg, Luke Miner, Erik Sverdrup, Stefan Wager, Marvin Wright, and Maintainer Julie Tibshirani. 2018. Package 'grf'.

Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. 2003. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678, New York City, NY.

Hal R Varian. 2016. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315.

Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the acl anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41.

Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Galen Weld, Peter West, Maria Glenski, David Arbour, Ryan Rossi, and Tim Althoff. 2020. Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference. *arXiv preprint arXiv:2009.09961*.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 4586. NIH Public Access.

Kui Yu, Jiuyong Li, and Lin Liu. 2016. A review on algorithms for constraint-based causal discovery. *arXiv preprint arXiv:1611.03977*.

A Appendix

We summarize the publication record encodings (treatments, outcomes, and covariates) used in our causal discovery and treatment effect estimation analyses in Table 4.

Treatment	Description
Publish on x	Binary encoding of whether x (one of our 50 research entities) was extracted from the publication's full text, using SpERT (Luan et al., 2018).
Scientist from c	Binary encoding of whether author is a scientist affiliated with country c (*i.e.,* author is associated with an insitution located in country c). There are six variations of this treatment, where c is one of the five countries wiht the greatest representation in ACL publications (the United States, China, Germany, Japan, France) or "Other".

Outcome	Description
Adopt x	Binary encoding of whether author will publish on x for the first time in the next calendar year.
Maintain x	Binary encoding of whether author previously published on x and has at least one publication on x in the following calendar year.
Retire/Pause x	Binary encoding of whether author previously published on x but has no publications on x in the following two calendar years.
Publication Increase Rate	The relative increase in the number of publications by author in the next calendar year, *i.e.,* $\frac{\#\ publications\ in\ year\ t+1}{\#\ publications\ in\ year\ t}$

Covariate	Description
# Papers	Total number of author's publications (cumulative since first publication).
# Papers (Last Year)	Number of author's publications within the last year.
# Papers (Last 5 Years)	Number of author's publications within the last five years.
# Co-authors	Number of co-authors linked to author in the collaboration network (by year).
Avg. # Papers Co-authored	Average paper count per co-author, *i.e.,* $\frac{\#Papers}{\#Co-Authors}$
Co-author Likelihood	Likelihood that a randomly selected co-author was a co-author on the publication, *i.e.,* $\frac{Avg.\#PapersCo-authored}{\#Papers}$
Centrality	Degree centrality of the author in the collaboration network, for the publication's calendar year.
Page Rank	Page rank of the author in the collaboration network, for the publication's calendar year.
Time Since First Paper	Number of years since the author's first publication in the ACL anthology.
Conference	Binary encoding of whether the paper is published in a conference (1) or journal (0).
# Authors	Publication's total number of authors.
Page Length	Page length of the publication's full text PDF.
# Words	Total number of words in the publication's full text.
Gunning Fog Index	Gunning Fog Index readability measure (Gunning et al., 1952) calculated using the publication text.

Table 4: Overview of Treatments (above), Outcomes (center), and Covariates (below).