Causal Inference from News Streams

Ananth Balashankar [1] **Sunandan Chakraborty** [2] **Lakshminarayan Subramanian** [1] **Samuel Fraiberger** [1 3]
Srikanth Jagabathula [1]

1. True Causality vs Predictive Causality

Causal inference is currently of immense interest in fields like medicine, economics and social science where interpretability of machine learning models is vital for increasing trust of practitioners. However, (Pearl, 2009) shows that true causal inference is not possible through post-facto observational studies alone due to the possibility of unobserved variables, which might be the underlying cause. However, there is still value in establishing a link probabilistically from an independent variable (X) to a dependent variable (Y) for guidance to understand the conditions in which they can be studied, denoted by $Pr(Y = y_0|X = x_0) >> Pr(Y = y_0|X = x_1)$. We adopt this view and extend this framework to extract links between various events occurring in the news.

We understand that determining causality is hard, but causality is a means to an end. In many online systems like search, recommendations and ad placement, A/B experiments are designed to confirm hypotheses initially intuited using predictive causality. Predictive causality in itself can prove to be quite useful if the problem to be solved is to improve causal experiment design. For example, using predictive causality, a surveyor can customize and reduce the number of questions and still capture all the required information in a survey. We study the problem of predictive causality in news streams and analyze its various use cases and the implications of solving such a problem.

2. Predictive Causality in News Streams

One such causality framework is the Granger causality framework. Given two time-series X and Y , the Granger causality test checks whether the X is more effective in predicting Y than using just Y and if this holds then the test concludes X Granger-causes Y (Granger et al., 2000; Granger, 2004; 2001). For news streams, we try to understand relationships between words that describe events to potentially uncover *hidden* relationships between news events, that may be well separated over time. For example, malaria epidemics are known to be triggered by monsoons in tropical climates; this hidden relationship may manifest in new streams with a frequency spike in the word "monsoon" followed by a frequency spike in the word "malaria", a few weeks later. Thus, mining large news datasets can potentially reveal influencing factors behind the surge of a particular word in news. This notion can be generalized to discover the influence of one event to another, where the events are manifested by specific words appearing in news. Many existing statistical methods that are used to determine relationships are *associative*. A drawback of these associative relationships (Newman et al., 2006; Nakashole et al., 2012; Light et al., 2008) is that they are solely based on the context and co-occurrence (Blei et al., 2003; Mikolov et al.). However, two causal events might appear across two articles that may not appear to be related or may be separated across time.

We construct a Word Influencer Network (WIN) by identifying Granger causal pairs of words, and combining them to form a network of words, where the directed edges depict potential influence between words. The network provides a more holistic view of the causal information flow by overcoming a common drawback of pair-wise Granger causality, when the true relationship involves three or more variables (Maziarz, 2015). WIN can offer the following that can significantly increase the benefits of using news for analytics – (1) Detection of influence path, (2) Discovery of unknown facts, (3) Hypothesis testing and feature extraction for experiment design.

3. Measuring Predictive Causality

Constructing a word network using an exhaustive set of word pairs to limit the problem of confounding variables can be computationally challenging and prohibitively expensive when the vocabulary size is fairly large. This is even after using a reduced set of informative words and only the collocated phrases, the vocabulary size is around $30,000$. One solution to this problem is considering the Lasso Granger method (Arnold et al., 2007) that applies regression to the neighborhood selection problem for any

*Equal contribution [1]New York University [2]Indiana University [3]World Bank. Correspondence to: Ananth Balashankar <ananth@nyu.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

word, given the fact that the best regressor for that variable with the least squared error will have non-zero coefficients only for the lagged variables in the neighborhood. The Lasso algorithm for linear regression is an incremental algorithm that embodies a method of variable selection. The output is w, which is obtained by minimizing the sum of the average squared error of regressing for y, and a constant times the L1-norm of the coefficients(Tibshirani, 1994).

We also note that using linear regression for such a Granger causal link detection is not a necessity. Complex non-linear links can be derived using neural networks as shown in (Tank et al., 2018). They model the task of Granger causality using component-wise multilayer perceptrons (cMLP) and cLSTMs. In the cMLP architecture proposed, they model each time series as a single MLP. In order to model long time lags in Granger causality, they use component-wise LSTMs for each time series. The hidden layers of the neural networks are then used as variables in the final linear regression equation to give the output variable. They penalize the non-zero weights in the input layer using the full Lasso penalty and the hierarchical group Lasso penalty and show that the group Lasso penalty chooses a suitable lag for each of the time series. The underlying basis of using component wise neural networks is to provide interpretability of the overall model, i.e if all the weights corresponding to a given time series' component is zero, then we can determine non-causality through it. However, this comes at the cost of overparametrization in neural networks.

3.1. Compact Causal Representations

One way to minimize the number of parameters to train would be to train a graph neural network (You et al., 2018; Li et al., 2018; Velikovi et al., 2018) which tries to perform the causal edge prediction task based on message passing in a subset of representative causal networks bootstrapped using the above approach. This approach could be especially relevant in our news causality task as there is redundancy in the nodes of the network. For example, two synonyms may be used interchangeably in the news and hence learning one set of causal edges would suffice to extend it for the other using semantic word embeddings (Peters et al., 2018; Sharp et al., 2016; Zhao et al., 2017). Or, in the case of granger causality, two words which have very similar time series (low Wasserstein distance) would have identical causal edges in the network (Courty et al., 2018). Thus, it is important to embed the nodes in the network into a low dimensional space which can capture both semantic and temporal similarity.

The number of nodes of WIN corresponds to the vocabulary size and it can be hard to visualize the graph due to its size. We make the graph coarser by reducing the nodes to *topics* learned from the news corpus using Latent Dirichlet Allocation (LDA)(Blei et al., 2003). Influence is generalized at the topic level by filtering edges based on the strength of inter-topic influence relationships measured by the total number of edges between nodes of two topics. Constructing such a WIN, can be useful for better experimental design and phenomenon understanding for social scientists. For example, the most important link we constructed on the Times of India news archive dataset was from topics "Politics" to "Crime".

4. Our Experiences

We quantitatively evaluate WIN by using it to extract features for stock price prediction and obtained two orders lower prediction error compared to a similar causal graph based method (Kang et al., 2017) which used time series of known cause-effect tuples. The dataset[1] contains news crawled from 2010 to 2013 using Google News APIs and New York Times data from 1989 to 2007 with 12,804 unigrams and 25,909 bigrams. We construct WIN from the time series representation of unigrams and bigrams, as well as the 10 stock prices[2] from 2013 we use for prediction. Qualitatively, we used WIN on the Times of India news archive from 2006 to 2015 to see if two events which are established to be causal in the news with an indicative verb like "caused" are linked in the network. We validated that 67% of the manually found causal evidence is linked through a directed edge and the rest 33% through a path of length 2 as shown in Table 1.

In (Chakraborty et al., 2016), our early experiences aimed at constructing a model to relate events in news streams to food price fluctuations in the Indian context. We demonstrated a high accuracy for spike prediction of price fluctuations which led to our question of predictive causality that can relate news events with price fluctuations.

Table 1. Causal Evidence in WIN with time lags (days)

Word pairs	**Words of the influence path**
land, budget	allot-land –(22)– railway-budget
price, land	price-hike –(12)– land
strike, law	terror-strike –(25)– law ministry
land, bill	land-reform –(25)– bill-pass

5. Discussion

We present WIN, a framework for building unsupervised predictive causal networks which capture hidden relationships between words in news streams. We demonstrate the power of these networks in evaluating causal hypotheses and extracting features for structured prediction tasks. Given its

[1]https://github.com/dykang/cgraph

[2]https://finance.yahoo.com

capacity to capture both temporal and semantic similarity, we intend to explore these predictive causal networks as building blocks for building complex reasoning systems.

References

Arnold, Andrew, Liu, Yan, and Abe, Naoki. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pp. 66–75, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-609-7. doi: 10.1145/1281192.1281203. URL http://doi.acm.org/10.1145/1281192.1281203.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

Chakraborty, Sunandan, Venkataraman, Ashwin, Jagabathula, Srikanth, and Subramanian, Lakshminarayanan. Predicting socio-economic indicators using news events. In *KDD 2016 - Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-August-2016, pp. 1455–1464. Association for Computing Machinery, 8 2016. doi: 10.1145/2939672.2939817.

Courty, Nicolas, Flamary, Rmi, and Ducoffe, Mlanie. Learning wasserstein embeddings. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJyEH91A-.

Granger, Clive W.J. Essays in econometrics. chapter Investigating Causal Relations by Econometric Models and Cross-spectral Methods, pp. 31–47. Harvard University Press, Cambridge, MA, USA, 2001. ISBN 0-521-79697-0. URL http://dl.acm.org/citation.cfm?id=781840.781842.

Granger, Clive W.J. Time series analysis, cointegration, and applications. *American Economic Review*, 94(3): 421–425, June 2004. doi: 10.1257/0002828041464669. URL http://www.aeaweb.org/articles?id=10.1257/0002828041464669.

Granger, Clive WJ, Huangb, Bwo-Nung, and Yang, Chin-Wei. A bivariate causality between stock prices and exchange rates: evidence from recent asianflu. *The Quarterly Review of Economics and Finance*, 40(3):337–354, 2000.

Kang, Dongyeop, Gangal, Varun, Lu, Ang, Chen, Zheng, and Hovy, Eduard. Detecting and explaining causes from text for a time series event. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2758–2767, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1292.

Li, Yujia, Vinyals, Oriol, Dyer, Chris, Pascanu, Razvan, and Battaglia, Peter. Learning deep generative models of graphs, 2018. URL https://openreview.net/forum?id=Hy1d-ebAb.

Light, Marc, Schilder, Frank, Kondadadi, Ravi Kumar, Dozier, Christopher C, Liao, Wenhui, and Veeramachaneni, Sriharsha. Systems, methods, and software for entity extraction and resolution coupled with event and relationship extraction, December 22 2008. US Patent App. 12/341,926.

Maziarz, Mariusz. A review of the granger-causality fallacy. *The Journal of Philosophical Economics*, 8(2):6, 2015. URL https://EconPapers.repec.org/RePEc:bus:jphile:v:8:y:2015:i:2:n:6.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Gregory S., and Dean, Jeffrey. Distributed representations of words and phrases and their compositionality. NIPS '13, pp. 3111–3119.

Nakashole, Ndapandula, Weikum, Gerhard, and Suchanek, Fabian. Patty: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1135–1145. Association for Computational Linguistics, 2012.

Newman, David, Chemudugunta, Chaitanya, Smyth, Padhraic, and Steyvers, Mark. Analyzing entities and topics in news articles using statistical topic models. In *ISI*, pp. 93–104. Springer, 2006.

Pearl, Judea. Causal inference in statistics: An overview. *Statistics Surveys*, 2009.

Peters, Matthew E., Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, and Zettlemoyer, Luke. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. URL http://arxiv.org/abs/1802.05365.

Sharp, Rebecca, Surdeanu, Mihai, Jansen, Peter, Clark, Peter, and Hammond, Michael. Creating causal embeddings for question answering with minimal supervision. *CoRR*, abs/1609.08097, 2016. URL http://arxiv.org/abs/1609.08097.

Tank, A., Covert, I., Foti, N., Shojaie, A., and Fox, E. Neural Granger Causality for Nonlinear Time Series. *ArXiv e-prints*, February 2018.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

Velikovi, Petar, Cucurull, Guillem, Casanova, Arantxa, Romero, Adriana, Li, Pietro, and Bengio, Yoshua. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

You, Jiaxuan, Ying, Rex, Ren, Xiang, Hamilton, William L., and Leskovec, Jure. Graphrnn: A deep generative model for graphs. *CoRR*, abs/1802.08773, 2018. URL http://arxiv.org/abs/1802.08773.

Zhao, Sendong, Wang, Quan, Massung, Sean, Qin, Bing, Liu, Ting, Wang, Bin, and Zhai, ChengXiang. Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pp. 335–344, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4675-7. doi: 10.1145/3018661.3018707. URL http://doi.acm.org/10.1145/3018661.3018707.