**Understanding Causality in NLP**

This paper attempts to understand the role of causality in Natural Language Processing (NLP) and explore how current NLP research utilizes concepts of causality. To begin, we must first understand why causality is important in today's machine learning landscape and how it bridges gaps in today's algorithms. From there, we can provide an example of how causality can be used to address problems by reviewing research in this area.

Machine learning is an integral aspect of computing in our world today. We build a mathematical model to predict an outcome of an event given some initial set of parameters and data. Its utility lies in its ability to progressively improve the performance of its prediction as it "learns" more about the problem. However, there are inherent limitations to this approach. Namely, models are limited to making predictions. If want to understand the cause of a particular pattern we see occurring in data, it is challenging. Even from a human perspective, it is extremely difficult to judge what causes what. For example, there is a correlation between temperature and pressure, but it is not clear if it is an increase in temperature that causes pressure to increase or if it is an increase in pressure that causes the temperature to increase.

Thus, it is much easier to develop a model that only strives to find a correlation between variables rather than understanding the causal relationship between them. But we must keep in mind that correlation does not imply causation. To better understand the importance of causality, let us take an example where we have a dataset that contains pictures of cows and camels [1]. We build a model that will tell us whether the picture has a cow in it. Surprisingly, we observe that our model has an accuracy of over 95%. However, if we break down how our algorithm characterizes these pictures, we see that the model is actually looking at the environment rather than the actual subject of the picture. Our dataset consisted of cow pictures taken on a green pasture. The camels were almost always were pictured in a dessert with a beige background. If we had tested our model on pictures where the locations of the animals were flipped (a picture of a cow taken on a sandy beach or a camel in an oasis), we would find that our model would fail to classify this dataset correctly. From this example, we can comprehend the importance of building a model that is based on features that will not change. This is the basis of causality. We strive to build relationships between features that will not vary with new incoming data and over time. This makes models built on this principle robust and more likely to be transferable between different, unknown environments. By doing so, we address several problems that have hindered the progress of machine learning so far: lack of robustness to unseen data, lack of algorithmic interpretability, inbuilt bias by using a biased/limited dataset, datasets that have skewed underlying distributions (i.e. healthcare-related datasets and datasets put together for fraud detection), etc. By utilizing elements of causality in current machine learning algorithms we can attempt to resolve these fundamental issues.

But that brings up the next question: How do we identify a causal relationship? Let us consider another example. Suppose we have a statement, "Covid causes coughing." We build a model that can identify the correlation that given we have Covid, we would cough. We can see that this correlation should not break no matter what changes. If the number of people who test positive increases, proportionally the number of people who are affected by coughing should also increase. On the other hand, if I propose that "Coughing causes Covid", we are identifying the probability that given I am coughing, how likely is it that I have Covid. However, we already know there are confounding factors. If it is allergy season, coughing will naturally increase. An increase in people coughing does not imply that the number of people who test positive also increases. Another implication is that if you address the problem of Covid, coughing will

naturally stop. But if you address the problem of the cough, the number of people who have Covid will not change. This distinction is important because when we are trying to build causal relationships, we do not want to find correlations where X correlated with Y, but Y does not correlate with X. We want to find causal relationships as these are relationships that will not change when new elements are introduced.

So how does establishing such a relationship enable us to build better natural language processing systems? Consider a text classification problem where we aim to identify Twitter tweets that contain misinformation. We have some cited source, A, that has been validated by experts. A causal relationship would be to use that dataset to check to see if that tweet contains misinformation. An anticausal, unstable, relationship would be to use the number of retweets to identify if the post has misinformation – there is evidence that a correlation exists between the propagation of misinformation and the number of retweets as found in the 2016 election [3].

To represent this relationship mathematically, consider a causal chain A → Y → Z [6] where A represents a dataset on which we want to train, it should be a stable source of information; Y represents the dataset we want to label, a collection of tweets; and Z represents an unstable dataset, the number of retweets for each tweet in Y. Our goal is to develop a feature set, I, that we can correlate to A and Z. The knowledge transfer from I → A and I → Z should be stable and robust to distribution changes.

And now we come to the crux of the issue. How do we identify these features that enable us to build such a relationship? This is a problem that many disciplines from Psychology to Philosophy to Computer Science have attempted to explain and have, so far, not yet reached a consensus. Nevertheless, we can look at current attempts in the field today that utilize causality.

One of the most popular NLP methods that utilize causality is to use "approaches that employ linguistic, syntactic, and semantic pattern matching" [2]. Our daily language naturally integrates words that denote causal relationships like "if… then", "because", "since", "cause", "effect", etc. Koo *et al.* [4] lists several explicit representations of such language: causal links which are words that link clauses (e.g. so, hence, therefore); causative verbs which are transitive (e.g. to cause to [verb]); resultative constructions where the object of a verb is followed by a description in a sentence (e.g. I filled a bucket with water); if-then conditionals; and causation adverbs and adjectives where descriptive words contain causal elements (e.g. fatal/fatally that can be paraphrased as "to cause to die"). However, the limitation of this approach is that it ignores the implicit representation of causality in text. Suppose we describe a particularly ferocious dog. The following sentence is about a boy who is running away from this dog. As a human reader, we can understand that the boy is probably running away because of fear and that the dog is chasing the boy either from hunger or aggravation. Unless the cause of why the boy is running away from the dog is explicitly stated in a representation that is described by Koo, it is unlikely that the model will recognize this connection. But if the domain of the text is known, we can construct specific representations for some common causal links.

Another limitation of the above method is that we are limited to knowledge within that domain or that particular text. Assume that we are reading an article about how the market value of a pharmaceutical company has increased. As a human, we may connect this back to the recent release of a new drug that can effectively treat cancer by that company. Unless the article explicitly mentions this connection, the model we generate will miss this causal link.

This leads us to a second approach where we use an additional feature, time, to derive causal relationships in a text. There are two general approaches to this method. One is to use infer causal relationships over a set of random variables that represent different events. We pair

events together and use A/B testing to find which set of samples and observations are more strongly correlated. The other approach is to infer causality over a time series. We can pair up another dataset like daily stock market prices with a series of articles over some time frame and determine whether a release of an article influenced the stock market or the stock market influenced a release of the article [5].

Another popular approach that is being explored is to use transfer learning. Transfer learning is the "ability of a model to generalize over previously unseen domains and/or tasks" [7]. Essentially, it is a machine learning method where we first generate one model for a particular task and reuse it for a second task. The second task is usually a different but related model. In the context of NLP, it would be similar to developing a model to analyze tweets from Twitter and applying the model to a new dataset, say Reddit posts. We would need to generalize our first model and make sure that we are not making assumptions or creating dependencies that will only be found in the Twitter dataset (e.g. limitation of the length of characters allowed in the tweet) so that we can also apply it to posts on Reddit. In other words, the features we use to analyze the first dataset should be features that will be found in both the Twitter and Reddit datasets. These features we choose should have a causal relationship with the outcome we are observing. A study conducted by Moeed *et al*. [7] showed the effectiveness of using transfer learning with a specific neural net (Progressive Neural Networks (PNNs)). They proved that it is possible to build a classifier that outperforms traditional PNNs with regards to natural language processing tasks (i.e. sentiment analysis).

Causality is a difficult subject to begin contemplating due to its inherently complex nature, but if you can successfully derive causal relationships, you can build a robust model against unseen data and is generalizable to multiple domains.

## Citations


[1] Arjovsky, Martin, et al. "Invariant risk minimization." *arXiv preprint arXiv:1907.02893* (2019).

[2] Asghar, Nabiha. "Automatic extraction of causal relations from natural language texts: a comprehensive survey." *arXiv preprint arXiv:1605.07895* (2016).

[3] Bovet, Alexandre, and Hernán A. Makse. "Influence of fake news in Twitter during the 2016 US presidential election." Nature communications 10.1 (2019): 1-14.

[4] C. S. Khoo, J. Kornfilt, R. N. Oddy, and S. H. Myaeng, "Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing," Literary and Linguistic Computing, vol. 13, no. 4, pp. 177–186, 1998.

[5] Kim, Hyun Duk, et al. "Mining causal topics in text data: iterative topic modeling with time series feedback." *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2013.

[6] Magliacane, Sara, Tom Claassen, and Joris M. Mooij. "Joint causal inference on observational and experimental datasets." *arXiv preprint arXiv:1611.10351* (2016).

[7] Moeed, Abdul, et al. "An Evaluation of Progressive Neural Networksfor Transfer Learning in Natural Language Processing." *Proceedings of The 12th Language Resources and Evaluation Conference*. 2020.

[8] Wood-Doughty, Zach et al. "Challenges of Using Text Classifiers for Causal Inference." *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* vol. 2018 (2018): 4586-4598.