Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Doctoral Consortium Track

On Building a Semi-Automated Framework for Generating Causal Bayesian Networks from Raw Text

Solat J. Sheikh
Artificial Intelligence Lab, Institute of Business Administration, Karachi, Pakistan
sjsheikh@iba.edu.pk

Abstract

The availability of a large amount of unstructured text has generated interest in utilizing it for future decision-making and developing strategies in various critical domains. Despite some progress, automatically generating accurate reasoning models from the raw text is still an active area of research. Furthermore, most proposed approaches focus on a specific domain. As such, their suggested transformation methods are usually unreliable when applied to other domains. This research aims to develop a framework, SCANER (Semi-automated CAusal Network Extraction from Raw text), to convert raw text into Causal Bayesian Networks (CBNs). The framework will then be employed in various domains to demonstrate its utilization as a decision-support tool. The preliminary experiments have focused on three domains: political narratives, food insecurity, and medical sciences. The future focus is on developing BNs from political narratives and modifying them through various methods to reduce the level of aggressiveness or extremity in the narratives without causing conflict among the masses or countries.

1 Introduction

The raw text constitutes a significant and rapidly growing form of data generated by various sources daily. The primary sources of raw text include social networking platforms (tweets and comments), hospitals (clinical notes), e-commerce websites (product reviews and ratings), political master narratives, news portals, and publications. This raw textual data holds valuable information helpful in making future decisions. However, the raw text accumulated from different sources is unstructured in nature, requiring extensive reading and processing to comprehend. It is, therefore, desirable to transform the raw text into diagrammatic models (such as knowledge graphs and causal Bayesian networks), which are easy to read and comprehend.

The causal reasoning mechanism of the generated diagrammatic models can be used for decision-making purposes. It is worth mentioning that causal reasoning is essential for decision-making in critical domains, such as health care, disaster management, theft detection, finance, and law. These domains directly impact human lives; hence, incorrect decisions might have substantial adverse impacts.

2 Related Work

The existing literature on generating causal Bayesian networks from the raw text can be divided into two categories: manual generation and automatic generation. Traditionally, causal models were built manually by subject matter experts. Because of this expert-driven approach, these models showed higher accuracy and were more trustworthy for decision-making in critical domains. One such example is the recent work by [Levis, 2019], who employed a series of human-guided steps to generate probabilistic causal models from political master narratives. However, it must be noted that manually building causal models from raw text is labor-intensive and time-consuming.

A few methods have recently been reported in the literature that utilize the recent advances in natural language processing (NLP) and overcome the limitations of the manual process. [Doan et al., 2019] used an automated approach to extract causality from health-related Twitter data. However, the proposed methodology used only three target effects and six syntactic patterns to detect the cause-effect relationships, thereby limiting the scope of their application. Later, [Sharp et al., 2019] developed an automated, rule-based system, Eidos, for identifying causality in raw text. However, the causal diagrams generated by Eidos were found to be less interpretable because of unlabeled edges. The models also showed limited coverage in domains other than food insecurity. [Min et al., 2021] have also mentioned this limitation of Eidos as it showed minimal coverage of the events related to COVID-19.

Despite these efforts, the automated process of causal model construction suffers from various challenges due to the linguistic complexity and ambiguity of the texts. In addition, the subject matter experts can only rely on such automatically generated models when they explain their suggested decisions in complicated scenarios and critical domains. Therefore, human involvement is essential for verifying the automatically generated models. Another area for improvement is that each system is designed specifically for a single domain

and shows limited or no performance on the text belonging to other domains. Moreover, a few approaches have also been reported that usually rely on numeric data that is either derived from the occurrence or frequency of causal relationships within the raw text or from the statements containing amounts in the form of percentages. Due to space limitations, not all such efforts have been cited here. One major drawback of these approaches is that they are only effective when sufficient data is available. However, this may not be the case in every domain. This further highlights the need for a semi-automated framework that is not dependent on having a sufficient amount of text for extracting probabilistic causal models from the raw text.

3 Research Objectives

This research aims to combine the strengths of both manual and automated methodologies for developing a generalized semi-automated framework for converting the raw text of any size and domain into Causal Bayesian Networks. It first generates a causal network by extracting the causal triples from raw text. Afterwards, the proposed framework would convert the causal network into Causal Bayesian Network by populating them with conditional probabilities (CPTs). This research will also demonstrate the utility of the resultant networks as decision-support tools by applying them specifically in the political domain, inspired by Levis [2019].

4 Proposed Methodology

The proposed approach, so far, has been divided into seven major components: manual simplification of text, preprocessing, triples' extraction, causal triples filtration, causal network generation, Causal Bayesian Network generation, and utilization of the framework as a decision support tool in the political domain. A unified application of these components will result in a generalized semi-automated framework for generating Causal Bayesian Networks from raw text. Figure 1 shows the workflow of the proposed methodology. As datasets with ground truth for validation are unavailable, we aim to assess the proposed framework by obtaining feedback from human annotators.

So far, the component of extracting causal networks (without CPTs) from raw text has been developed. A prototype has been deployed on the streamlit cloud for visualization. It must be noted that the website[1] is a work in progress and will continue to evolve. The approach has been tested using the text from three domains (political narratives, food insecurity, and medical science). In addition, generated causal models were compared against the ones produced by Eidos, Sharp et al. [2019], one of the most popular causality extraction systems in the literature. The comparison demonstrates the advantages of the proposed approach in generating dense and accurate causal networks from raw text. The findings have been submitted and are under review in an impact factor journal.

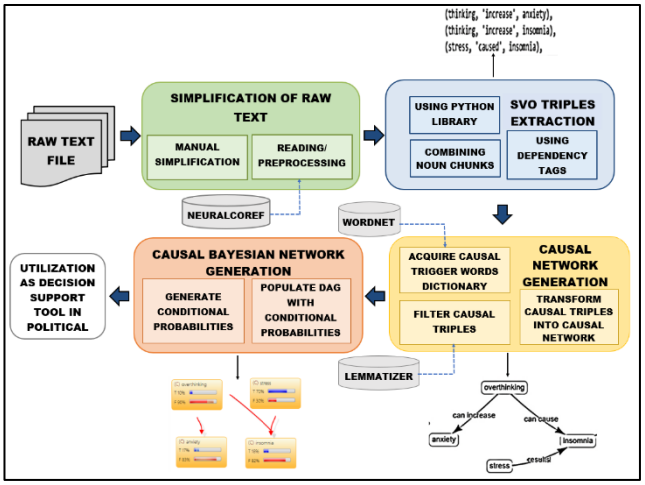

Figure 1: Structure of the Framework (SCANER) for Constructing Causal Bayesian Networks

5 Conclusion and Future Work

This research aims to develop a framework that transforms a given raw text into Causal Bayesian Networks without relying on the sufficiency of the text for determining CPTs.

My current focus is on converting causal networks into Causal Bayesian Networks by populating them with CPTs. In addition, I am exploring ways to assess alternative courses of action by modifying the generated models, especially for text containing political narratives.

References

[Doan et al., 2019] Doan, Son, Elly W Yang, Sameer S Tilak, Peter W Li, Daniel S Zisook, and Manabu Torii. 2019. "Ex-tracting Health-Related Causality from Twitter Messages Us-ing Natural Language Processing." BMC Medical Informatics and Decision Making 19 (3): 71–77.

[Levis, 2019] Levis, Alexander H. 2019. "On Narrative Model-ing And Assessment For Strategic Change." In ECMS, 393–99.

[Min et al., 2021] Min, Bonan, Benjamin Rozonoyer, Haoling Qiu, Alexan-der Zamanian, and Jessica MacBride. 2021. "Excava-torcovid: Extracting Events and Relations from Text Corpora for Temporal and Causal Analysis for Covid-19." ArXiv Preprint ArXiv:2105.01819.

[Sharp et al., 2019] Sharp, Rebecca, Adarsh Pyarelal, Benjamin Gyori, Keith Alcock, Egoitz Laparra, Marco A Valenzuela-Escárcega, Ajay Nagesh, et al. 2019. "Eidos, INDRA, & Delphi: From Free Text to Executable Causal Models." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations).

[1] https://solatjabeen-causal-graph-acquisition-streamlitproject-2k94yr.streamlitapp.com/