Learning Causal Bayesian Networks from Text

Farhad Moghimifar, Afshin Rahimi, Mahsa Baktashmotlagh and **Xue Li**
The School of ITEE
The University of Queensland, Australia
{f.moghimifar,a.rahimi,m.baktashmotlagh}@uq.edu.au
xueli@itee.uq.edu.au

Abstract

Causal relationships form the basis for reasoning and decision-making in Artificial Intelligence systems. To exploit the large volume of textual data available today, the automatic discovery of causal relationships from text has emerged as a significant challenge in recent years. Existing approaches in this realm are limited to the extraction of low-level relations among individual events. To overcome the limitations of the existing approaches, in this paper, we propose a method for automatic inference of causal relationships from human written language at conceptual level. To this end, we leverage the characteristics of hierarchy of concepts and linguistic variables created from text, and represent the extracted causal relationships in the form of a Causal Bayesian Network. Our experiments demonstrate superiority of our approach over the existing approaches in inferring complex causal reasoning from the text.

1 Introduction

Causation is a powerful psychological tool for human to choreograph his surrounding environment into a mental model, and use it for reasoning and decision-making. However, inability to identify causality is known to be one of the drawbacks of current Artificial Intelligence (AI) systems (Lake et al., 2015). Extraction of causal relations from text is necessity in many NLP tasks such as question answering and textual inference, and has attracted a considerable research in recent years (Wood-Doughty et al., 2018; Zhao et al., 2018, 2017; Ning et al., 2018; Rojas-Carulla et al., 2017). However, the state-of-the-art methods are limited to the identification of causal relations between low-level individual events (Dunietz et al., 2017; Hidey and McKeown, 2016; Mirza and Tonelli, 2016) and fail to capture such relationships at conceptual level. Furthermore, relying on linguistic features limits the identification of causal relations to those whose cause and effect are located in the same sentence or in consecutive sentences.

In this paper, we propose a method for extracting concepts and their underlying causal relations from written language. Furthermore, to leverage the extracted causal information, we represent the causal knowledge in the form of a **C**ausal **B**ayesian **N**etwork (CBN). Having this tool enables answering complex causal and counter-factual questions, such as: *How psychotherapy can affect the patient's emotion?*, or *What would happen if instead of medicine X, medicine Y was prescribed?*

The contribution of this paper is three-fold. Firstly, we focus on identifying causal relation between concepts (e.g. *physical activity* and *health*). Secondly, We propose a novel method to represent the extracted causal knowledge in the form of a Causal Bayesian Network, enabling easy incorporation of this invaluable knowledge into downstream NLP tasks. Thirdly, we release PSYCAUS dataset which can be used to evaluate causal relation extraction models in the domain of psychology [1]. In addition, our proposed method identifies causality between concepts independent of their locations in text, and is able to identify bi-directional causal relations between concepts, where two concepts have causal effect on each other. By aggregating linguistic variable, we construct a hierarchy where each variable, e.g. delusional disorder, lies under its related concept, e.g. disorder. This hierarchical and inheritance structure allows for the inference of causal relations between concepts that are not directly discussed in the text.

In order to evaluate our proposed method, we gathered a corpus of psychological articles The experimental results shows that the proposed method performs significantly better than the state-of-the-

---

[1] https://github.com/farhadmfar/psycaus

art methods.

2 Related Works

Identification of causality in NLP is not trivial as a result of language ambiguity. Hence, most current approaches focus on verb-verb, verb-noun, and noun-noun relations. The explicit relations are often captured with narrow syntactic and semantic constructions (Do et al., 2011; Hendrickx et al., 2009; Mirza and Tonelli, 2016; Hidey and McKeown, 2016) which limits their recall. To go beyond surface form constructions few works have proposed neural models (Martínez-Cámara et al., 2017; Dasgupta et al., 2018; Zhao et al., 2017) covering wider causal constructions. However, most works don't go beyond extracting causality between adjacent events, and so lack the ability to capture causality in non-adjacent concept level, e.g. *genetics* and *hallucination*. Therefore, in this paper we propose a model for identifying causality between concepts, independent of their location, and represent the causal knowledge in form of a Causal Bayesian Network.

3 Methodology

Given the input, in form of human written language, we aim to extract the causal relation between concepts and represent the output in form of a Causal Bayesian Network. Hence, we split this task into three sub-tasks: extracting linguistic variables and values, identifying causality between extracted variables, and creating conditional probability table for each variable. In the following sub-sections each of these sub-tasks are explained.

3.1 Linguistic Variables

A *linguistic variable* is a variable which values are words in natural language (Zadeh, 1983). For example, if we consider the word Age as a linguistic variable, rather than having numeric values, its values are linguistic, such as young and old. A linguistic word is specified as $(C, T(C))$ where C in the name which represents the set of words or in other word the variable's name, and $T(C)$ is the set of represented words or linguistic values. In this context, a variable and corresponding value have an asymmetric relation, in which the hypernym (superordinate) implies the hyponym (subordinate).

In order to create a Bayesian Network (BN) from text, we first need to extract linguistic variables and values from our corpus. To this end, we leverage a probabilistic method introduced by Wu et al. (2012) to extract all possible *IsA* relations from corpus.

To enhance the accuracy of causality identification and runtime performance of our model, using Formal Concept Analysis (Ganter and Wille, 2012), we represent the extracted hypernym-hyponym relations in form of a hierarchy. In the context of our corpus, let V be the set of linguistic variables and v be the set linguistic values, we call the triple of (V, v, I) a formal context where V and v are non-empty sets and $I \subseteq V \times v$ is the incidence of the context. The pair of (V_i, v_i) is called a formal concept of (V, v, I) if $V_i \subseteq V, v_i \subseteq v, V_i^{'} = v_i$ and $v_i^{'} = V_i$, where $V_i^{'}$ and $v_i^{'}$ are sets of all attributes common to V and v, respectively. The formal concept of a given context are naturally in form of a subconcept-superconcept relation, given for formal concepts of (V_i, v_i) and (V_j, v_j) of (V, v, I) : $(V_i, v_i) \leq (V_2, v_2) \iff V_i \subseteq V_j(\iff v_i \subseteq v_j))$. Consequently, we can identify that every attributes in the former formal concepts are also in the latter. Hence, this set of formula gives us the hierarchy of superconcept-subconcepts. Since every link in this hierarchy implies inheritance, attributes at the higher level are inherited by lower nodes. Therefore, if a concept V_i at level n of our hierarchy has a causal relation with another concept V_j, all the subconcepts of V_i at lower level m (where $m < n$), also have causal relation with V_j.

3.2 Identifying Causality

The core part of this paper is to identify the cause-effect relation between concepts, or linguistic variables. In a lower-level approach, causality is usually presented by syntactic relations, where a word or a set of words implies existence of causality. For example, 'cause' in 'Slower-acting drugs, like fluoxetine, may cause discontinuation symptoms" indicates a causal relation. These set linguistic features can be shown either in form of a verb or a discourse relation. The Penn Discourse Tree Bank (PDTB) (Prasad et al., 2008) contains four coarse-grained relations, comparison, contingency, expansion and temporal, in which contingency may indicate causality. There are 28 explicitly causal marker out of 102 in PDTB, with the barrier of causal relation. Furthermore, we leverage the sets of verbs included in `AltLexes`, such as 'force' and 'caused', which show causality in a sentence. Using both discourse and verb makers of causality,

we create a database of cause-effect (Γ) from given sentences. To this end, each of the input sentences are split into simpler version, using dependency parser, and once any of causality markers are identified in a sentence, the stopwords from cause and effect parts are eliminated and the remaining words are stemmed. Having the constructed cause-effect database (Γ), the causal relation between two concepts is defined as:

$$\text{CR}(V_m, V_n) = \frac{\sum_{i=1}^{|V_m|} \sum_{j=1}^{|V_n|} \mathbb{1}[(v_m^i, v_n^j) \in \Gamma] \vec{v}_m^i \cdot \vec{V}_m}{\sum_{i=1}^{|V_m|} \vec{v}_m^i \cdot \vec{V}_m}$$

$$- \frac{\sum_{j=1}^{|V_n|} \sum_{i=1}^{|V_m|} \mathbb{1}[(v_n^j, v_m^i) \in \Gamma] \vec{v}_n^j \cdot \vec{V}_n}{\sum_{j=1}^{|V_n|} \vec{v}_n^j \cdot \vec{V}_n} \qquad (1)$$

where V_m and V_n are two concepts or linguistic variables in the concept space V, and v_m^i is the i-th value of V_m; the functions r and w are defined as below:

$$r(a, b) = \begin{cases} 1 & if (a, b) \in \Gamma \\ 0 & if (a, b) \notin \Gamma \end{cases} \qquad (2)$$

and

$$w(a, b) = 1 - S_c(a, b) = 1 - sim(a, b) \qquad (3)$$

where sim is Cosine similarity of words a and b. The purpose of function w is to measure the relevance of the value to the corresponding variable, in order to increase the influence of more relevant values. The output of CR can be categorised as follow:

$$CR(A, B) \in \begin{cases} (\mu, 1] & \text{A cause; B effect} \\ [-\mu, \mu] & \text{no causal relationship} \\ [-1, -\mu) & \text{B cause; A effect} \end{cases} \qquad (4)$$

where μ is a threshold given to the model as a hyper-parameter.

3.3 Creating Conditional Probability Distribution

Conditional Probability Distribution (CPD) shows conditional probability of corresponding values to a variables with respect to values of parents of the variable $P(X_i|Parents(X_i))$. In order to extend the implementation of CPD for sets of linguistic variables we use *Normalized Pointwise Mutual Information (PMI)* score to calculate the probability distribution (Bouma, 2009).

$$i_n(x, y) = (ln \frac{p(x, y)}{p(x)p(y)}) / - ln(p(x, y)) \qquad (5)$$

The reason behind using *PMI* comes from Suppes' Probabilistic theory of Causality (Suppes, 1973), where he mentions that possibility of an effect to co-happen with cause is higher than happening by itself. In mathematical word it can be shown

as $P(effect|cause) > P(effect)$, which can be easily written as $\frac{P(\text{cause, effect})}{P(\text{cause})P(\text{effect})} > 1$, similar to *PMI* for positive values.

To create a Causal Bayesian Network from textual data, let G be our graphical model, and $V = \{V_1, V_2, ..., V_n\}$ be the set of extracted linguistic variables (as defined in §3.1) from our corpus ζ. We define $Pa_{V_i}^G = \{Vj : Vj \xrightarrow{causal} V_i\}$, indicating set of nodes in ζ which have causal relation with V_i. By expressing P as:

$$P(V_1, V_2, ..., V_n) = \prod_{i=1}^{n} P(V_i|Pa_{V_i}^G) \qquad (6)$$

we can argue that P factorises over G. The individual factor $P(V_i|Pa_{V_i}^G)$ is called conditional probability distribution (explained in 3.3). In a more formal way, we define Causal Bayesian Network over our corpus ζ as $\beta = (G, P)$ where P is set of conditional probability distributions. In addition to the aforementioned usages of a CBN, having a Causal Bayesian Networks enables the possibility of answering questions in three different layers of Association, Intervention and Counter-factual (Pearl and Mackenzie, 2018).

4 Experimental Results

In this section, we evaluate the effectiveness of our method and compare it with that of the state-of-the-art methods on a collection of Wikipedia articles related to Psychology. Each article within this collection is selected based on the terms in APA dictionary of Psychology (VandenBos, 2007). This collection contains a total number of 3,763,859 sentences. Among all possible relation between concepts, we studied 300 relationships between concepts, which were annotated by 5 graduate student from school of Psychology. Each tuple of relationship in form of (A, B), were labelled as $-1, 0$, or 1, where 1 indicates $A \xrightarrow{cause} B$, -1 shows that $B \xrightarrow{cause} A$, and 0 implies no causal relations. With the overlap tuples (25%) we measured the degree of agreement between annotators using Fleiss' Kappa measure (Fleiss and Cohen, 1973), which was around 0.67. This indicates the reliability of our test setup.

We compare our model to the feature-based and distribution-based methods proposed by Rojas-Carulla et al. (2017), with different proxy projection functions, including { *w2vii, w2vio, w2voi, counts, prec-counts, pmi, prec-pmi*}. Furthermore, we compare our model to heuristic models,
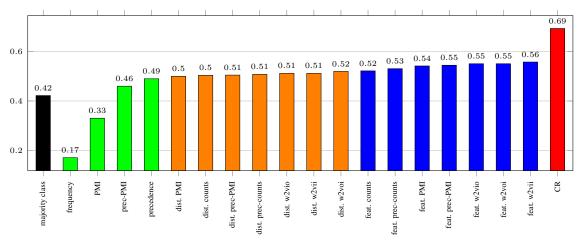
Figure 1: The test accuracy of *CR* compared with feature-based, distribution-based, heuristic methods, and the majority class.

consisting of frequency, precedence, PMI, PMI (precedence), where in each of the models two parameters are calculated, $S_{V_i \to V_j}$ and $S_{V_i \to V_j}$, indicating $V_i \xrightarrow{cause} V_j$ if $S_{V_i \to V_j} > S_{V_i \to V_j}$ and $V_i \xleftarrow{cause} V_j$ if $S_{V_i \to V_j} < S_{V_i \to V_j}$.

Figure 1 shows the accuracy of different methods for identifying causal relationships. We observe that our method (the red bar) outperforms all other approaches with an accuracy of 0.694. This indicates an improvement of 13% over the state-of-the-art feature-based methods (the blue bars), 17% over the distribution-based approaches (the orange bars), and 20% over the baseline methods (green bars). The baseline methods represented the worst performance, however, the accuracy achieved by precedence suggests that most of our corpus is written in form of active rather than passive voice, resulting in consequential connection between concepts.

To analyse the sensitivity of our method to the threshold μ in Equation 1, we trained the model on PSYCAUS's training set(D_{tr}), and analysed the development set performance in terms of macro-averaged F1 with a range of values $[0, 1)$ for μ. As shown in Figure 2, F1 score reaches the maximum of 0.66 with $\mu = 0.05$, well above a random classifier.

During the annotation process, we noticed that some concepts, e.g. *eating disorder* and *emotion*, may have bi-directional causal relations, depending on the context (eating disorder $\xleftrightarrow{cause}$ emotion). We ran our model against these examples and found out that our approach is interestingly capable of identifying these relations as well. In Equation 1, approximation of absolute values of both operands in the negation to one indicates bi-directional causality. While a bi-directional causal

relation cannot be presented in a CBN, as it is a *directed* graphical model, a decision tree can contain these types of information. In addition, some concepts, e.g. *delusional disorder* and *displeasure*, that have not been connected with any type of causality connectives were also accurately identified as causal relations. This is due to the hierarchical design of variable-values in our model.
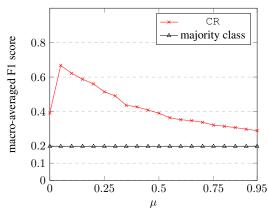

Figure 2: The macro-averaged F1 score of our proposed method on the development of PSYCAUS with different values of μ (Equation 1), compared with macro-averaged F1 score of the majority class model.

5 Conclusion

In this paper we have presented a novel approach for identifying causal relationship between concepts. This approach enables machines to extract causality even between non-adjacent concepts. Hence, a significant improvement was delivered comparing to naive baselines. Furthermore, we represented the causal knowledge extracted from human-written language in form of Causal Bayesian Network. To the best of our knowledge this representation is novel. Having a Causal Bayesian Network can empower many downstream applications, including question-answering and rea-

soning. Among all applications, causal and counter-factual reasoning, which can be build on top of the outcome of this paper, may address some current hallmarks of Artificial Intelligence systems.

References

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 306–316.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. Automatically tagging constructions of causation and their slot-fillers. *Transactions of the Association for Computational Linguistics*, 5:117–133.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Bernhard Ganter and Rudolf Wille. 2012. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1424–1433.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Eugenio Martínez-Cámara, Vered Shwartz, Iryna Gurevych, and Ido Dagan. 2017. Neural disambiguation of causal lexical markers based on context.
In *12th International Conference on Computational Semantics Short papers (IWCS)*.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2278–2288.

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

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Mateo Rojas-Carulla, Marco Baroni, and David Lopez-Paz. 2017. Causal discovery using proxy variables. *arXiv preprint arXiv:1702.07306*.

Patrick Suppes. 1973. A probabilistic theory of causality.

Gary R VandenBos. 2007. *APA dictionary of psychology*. American Psychological Association.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. *arXiv preprint arXiv:1810.00956*.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492. ACM.

Lotfi A Zadeh. 1983. Linguistic variables, approximate reasoning and dispositions. *Medical Informatics*, 8(3):173–186.

Sendong Zhao, Meng Jiang, Ming Liu, Bing Qin, and Ting Liu. 2018. Causaltriad: Toward pseudo causal relation discovery and hypotheses generation from medical text data. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 184–193. ACM.

Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, Ting Liu, Bin Wang, and ChengXiang Zhai. 2017. Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 335–344. ACM.