To Know the Causes of Things: Text Mining for Causal Relations

Fiona Anting Tan

Institute of Data Science, National University of Singapore
tan.f@u.nus.edu

Abstract

Causality expresses the relation between two arguments, one of which represents the cause and the other the effect (or consequence). Causal text mining refers to the extraction and usage of causal information from text. Given an input sequence, we are interested to know if and where causal information occurs. My research is focused on the end-to-end challenges of causal text mining. This involves extracting, representing, and applying causal knowledge from unstructured text. The corresponding research questions are: (1) How to extract causal information from unstructured text effectively? (2) How to represent extracted causal relationships in a graph that is interpretable and useful for some application? (3) How can we capitalize on extracted causal knowledge for downstream tasks? What tasks or fields will benefit from such knowledge? In this paper, I outline past and on-going works, and highlight future research challenges.

Introduction

Causal text mining refers to the extraction and usage of causal information from text. Given an input sequence, we are interested to know if and where causal information occurs. Our research is focused on the end-to-end challenges of causal text mining. This involves extracting, representing, and applying causal knowledge from unstructured text. The corresponding research questions are:

1. **Extraction:** How to extract causal information from unstructured text effectively?
2. **Representation:** How to represent extracted causal relationships in a graph that is interpretable and useful for some application?
3. **Application:** How can we capitalize on extracted causal knowledge for downstream tasks? What tasks or fields will benefit from such knowledge?

Extraction of Causal Relations

Our research focuses on enabling and designing such state-of-the-art solutions using transformers and pre-trained language models that demonstrated the effectiveness of extracting causal information from text.

Corpus Creation: Researchers need to have high quality and sufficient amount of training and testing data. Therefore, a key part of our research thus far works on corpus and repository creation.

First, we created the Causal News Corpus (CNC) (Tan et al. 2022, 2023a), a comprehensive corpus annotated of 3,767 sentences, of which, 1,982 are causal sentences that contain a total of 2,754 causal relations. We reported baseline experiments on two natural language tasks: (I) Causal Sentence Classification, and (II) Cause-Effect-Signal Span Detection.

Second, we recognize that some causal text mining data already exists in the field. However, there are inconsistencies across the datasets in terms of their text source, sentence lengths, linguistic construction and argument types. Thus, we proposed UniCausal (Tan, Zuo, and Ng 2023), a unified benchmark for causal text mining across three tasks: (I) Causal Sequence Classification, (II) Cause-Effect Span Detection and (III) Causal Pair Classification. We consolidated and aligned annotations of six high quality corpora, resulting in a total of 58,720, 12,144 and 69,165 examples for each task respectively.

Modelling: Harnessed with suitable datasets, we worked towards designing effective causal text mining models.

In the official run of FinCausal 2021 (Mariko et al. 2021), the model we designed obtained Precision, Recall, and F1 scores of 95.56%, 95.56% and 95.57% that all clinched 1^{st} place (Tan and Ng 2021). Exploiting the observation that words are more connected to other words with the same cause-effect type in a dependency tree, we incorporated dependency relation features through a graph neural network to construct useful graph embeddings to achieve better model performance.

To motivate continued interest in causal text mining modelling, we also organized two iterations of the Event Causality Identification Shared Task using CNC.

Beyond improving performance for causal text mining tasks, we also dived deeper into specific challenges of causal text mining. For example, we found that models misclassify on augmented sentences that have been negated or strengthened with respect to its causal meaning (Tan et al. 2021). This is worrying since minor linguistic differences in causal sentences can have disparate meanings. Therefore, we pro-

posed an algorithm to generate counterfactual causal sentences that serves as a contrast set to be included into the train set. By including a mixture of edits when training, we achieved performance improvements beyond the baseline across both models, and within and out of corpus' domain.

Representation & Application of Causal Relations

In our research, we are exploring two approaches to store and represent causal knowledge: (1) We could store causal relations in a knowledge graph (KG). Consider a causal graph $G = (V, E)$, where the nodes V are the objects in our universe, and the directed edges E represent the presence of causality between two nodes, where a Cause node points to an Effect node. (2) We could store causal knowledge latently within language models, especially if they were fine-tuned on a causal text mining task. We believe the approach depends on the application of the causal knowledge.

We have a few on-going projects working with causal KGs. In a recent publication (Tan et al. 2023b), we proposed a methodology to construct causal KGs from news using two steps: (1) Extraction of Causal Relations, and (2) Argument Clustering and Representation into KG. For extraction, we used a hybrid of BERT-based extraction models alongside pattern-based ones. For clustering, we utilized a topic modelling approach to cluster our arguments, so as to increase the connectivity of our graph. As a result, instead of 15,686 disconnected subgraphs, we were able to obtain 1 connected graph that enables users to infer more causal relationships from. Our final KG effectively captures and conveys causal relationships, validated through experiments, multiple use cases and user feedback. We are also working on automatic confounder detection from causal KGs (Tan and Ng 2023).

Future Work

We worked on multiple projects related to causal text mining, from extracting to representing and applying. We believe in the potential of harnessing causal relations in text for various real-world applications. However, we acknowledge that there are several challenges, of which three key issues are: (1) Understanding not only that a causal connection exists but also its degree of influence and the direction of causation is paramount in many applications. Addressing this challenge will require the development of more nuanced techniques and models that can assign weights and polarities to causal relations in a text. (2) Pertaining to the generalizability of causal relations, we noticed that clustering of co-referent arguments depends on the application. Balancing the need to generalize to uncover new relations with the risk of over-generalization is a complex task. Moreover, transitivity does not hold if the causal relations are specific to a situation or context. The violation of the transitivity property will violate many assumptions in a typical causal KG. This topic also has implications on dealing with few-shot and no-shot causal relations. (3) Finally, our current work has focused on the extraction of causal relations from reliable sources. Fake news identification is a large and growing research field, and potentially has many solutions we can adapt and learn from to identify wrong or subjective causal relations that might pollute a causal knowledge base. We hope to better analyze the problem and develop solutions by interacting with interdisciplinary AI researchers.

References

Mariko, D.; Akl, H. A.; Labidurie, E.; Durfort, S.; de Mazancourt, H.; and El-Haj, M. 2021. The Financial Document Causality Detection Shared Task (FinCausal 2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, 58–60. Lancaster, United Kingdom: Association for Computational Linguistics.

Tan, F. A.; Hazarika, D.; Ng, S.-K.; Poria, S.; and Zimmermann, R. 2021. Causal Augmentation for Causal Sentence Classification. In *Proceedings of the First Workshop on Causal Inference and NLP*, 1–20. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Tan, F. A.; Hettiarachchi, H.; Hürriyetoğlu, A.; Oostdijk, N.; Caselli, T.; Nomoto, T.; Uca, O.; Liza, F. F.; and Ng, S.-K. 2023a. RECESS: Resource for Extracting Cause, Effect, and Signal Spans. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. Bali, Indonesia: Association for Computational Linguistics.

Tan, F. A.; Hürriyetoğlu, A.; Caselli, T.; Oostdijk, N.; Nomoto, T.; Hettiarachchi, H.; Ameer, I.; Uca, O.; Liza, F. F.; and Hu, T. 2022. The Causal News Corpus: Annotating Causal Relations in Event Sentences from News. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2298–2310. Marseille, France: European Language Resources Association.

Tan, F. A.; and Ng, S.-K. 2021. NUS-IDS at FinCausal 2021: Dependency Tree in Graph Neural Network for Better Cause-Effect Span Detection. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, 37–43. Lancaster, United Kingdom: Association for Computational Linguistics.

Tan, F. A.; and Ng, S.-K. 2023. Economics Assistant for Robustness Checks (EconARC): Identifying Confounders from Causal Knowledge Graphs. In *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, November 6-10, 2023, Proceedings*, Lecture Notes in Computer Science. Springer.

Tan, F. A.; Paul, D.; Yamaura, S.; Koji, M.; and Ng, S.-K. 2023b. Constructing and Interpreting Causal Knowledge Graphs from News. In *Proceedings of the AAAI Summer Symposium Series (SuSS-23), Artificial Intelligence for FinTech (AI4FinTech)*. Singapore: Association for the Advancement of Artificial Intelligence.

Tan, F. A.; Zuo, X.; and Ng, S.-K. 2023. UniCausal: Unified Benchmark and Repository for Causal Text Mining. In Wrembel, R.; Gamper, J.; Kotsis, G.; Tjoa, A. M.; and Khalil, I., eds., *Big Data Analytics and Knowledge Discovery*, 248–262. Cham: Springer Nature Switzerland. ISBN 978-3-031-39831-5.