Causal Knowledge Extraction from Text using Natural Language Inference (Student Abstract)

Manik Bhandari[1], **Mark Feblowitz**[2], **Oktie Hassanzadeh**[2], **Kavitha Srinivas**[2], **Shirin Sohrabi**[2]
[1] Carnegie Mellon University
[2] IBM Research
mbhandar@cs.cmu.edu, mfeb@us.ibm.com, hassanzadeh@us.ibm.com,
kavitha.srinivas@ibm.com, ssohrab@us.ibm.com

Abstract

In this paper, we address the problem of extracting causal knowledge from text documents in a weakly supervised manner. We target use cases in decision support and risk management, where causes and effects are general phrases without any constraints. We present a method called CaKNowLI which only takes as input the text corpus and extracts a high-quality collection of cause-effect pairs in an automated way. We approach this problem using state-of-the-art natural language understanding techniques based on pre-trained neural models for Natural Language Inference (NLI). Finally, we evaluate the proposed method on existing and new benchmark data sets.

1 Introduction

Extracting causal knowledge from natural language descriptions of such knowledge in text documents is a challenging problem with a wide range of applications in AI systems. One major application area has been event forecasting (Radinsky, Davidovich, and Markovitch 2012), as well as decision support and risk management (Hassanzadeh et al. 2019, 2020). Our work targets this application area, where causes and effects are general phrases which may or may not be describing actions or events. A major challenge in applying state-of-the-art supervised knowledge extraction methods is the need for a large manually-annotated corpus, which is not feasible for large-scale generic causal knowledge extraction. Our focus in this paper is on weakly supervised methods where the input is a corpus of text documents that contain descriptions of causal knowledge required in the target application, and the output is a high-quality collection of cause-effect pairs, which can then be further processed, represented as a causal knowledge graph, and used as input for decision support or predictive analytics. Table 1 shows an example of a few cause-effect pairs extracted by one of our methods in an unsupervised way where the only input is a collection of Wikipedia articles about COVID-19.

2 CaKNowLI

Our approach for extracting causal relationships from text involves three stages described below - pattern matching, phrase extraction and natural language inference.

Cause	Effect
COVID-19 pandemic	wave of solidarity
COVID-19 pandemic	sharp increase in the use of telemedical services
COVID-19 outbreak	fear of a potential economic breakdown
COVID-19	reductions in bus route frequencies
fears of supply shortages	panic buying
panic buying	shortages of some products

Table 1: Examples of Cause-Effect pairs extracted by one of our proposed methods where the only input is a collection of Wikipedia articles on COVID-19.

Pattern Matching We start by creating a large list of nearly 200 causal patterns e.g. *X causes Y* [1] which is a subset of patterns used in Dunietz, Levin, and Carbonell (2017). We lemmatize all the patterns and the sentences to enable matching verbs in their root form and convert the patterns to regexes e.g. *"(.*) cause (.*)"* to match them against the sentence, obtaining the parts of the sentence corresponding to the cause and the effect.

Phrase Extraction Then, we extract phrases from the candidates and pair all combinations of causes with effects to form candidate cause-effect pairs. To extract phrases, we experiment with two phrase extraction techniques: (1) `NP` (Handler et al. 2016) which is a recently proposed algorithm for extracting noun phrases from sentences using Finite State Transducers, and (2) `CP` that is based on constituency parse of sentences to extract all kinds of phrases and not just noun phrases.

NLI Finally, we classify the obtained candidate cause-effect pairs as causal or non-causal using natural language inference. Let S_1 be the original sentence and (X, Y) be a candidate cause-effect pair. We construct a new causal sentence $S_2^i, i \in 1 \ldots k$ in k different ways based on $k = 7$

[1] Please see the supplementary material for a full list of causal patterns

syntactically different causal patterns. For instance, S_2^1 = *"X causes Y"*, S_2^2 = *"X is the reason for Y"*. We then use a pre-trained NLI model to get the probability P_i of inferring the causal sentence S_2^i from the original sentence S_1. We use the mean of the k probabilities as the probability of (X, Y) being causal.

3 Evaluation

We benchmark the performance of the proposed methods on two datasets - (1) the BECauSE 2.0 corpus (Dunietz, Levin, and Carbonell 2017) which consists of general phrases as causes and effects, tagged by annotators from within a sentence. And (2) the SemEval dataset, popularly used in prior causal extraction work (Sharp et al. 2016; Hassanzadeh et al. 2019), consisting of causal and non-causal relationships between words. We evaluate the performance of the proposed methods by matching the extracted causal pairs with the true pairs present in the dataset. An extracted pair is matched with a true pair if both its cause and effect are matched with those of the true pair. For the BECauSE dataset, two phrases are matched if their Jaccard Index is greater than 0.5. However, since the SemEval dataset consists only of words as causes and effects, we check if the true word from the dataset is *contained* within the extracted phrase.

4 Experiments

Please see the supplementary material for a detailed description of the hyperparamters.

4.1 Overall Results

In Table 2 we show the performance of our models. We can observe that if we provide the context as well as the true cause and effect to the `NLI` model, its performance increases significantly. We also observe that phrase extraction using the constituency parse (`CP`) performs better than NPFST (`NP`). However, the `PM+NLI` approach is better than phrase extraction based models. This is likely due to the fact that for short, well formed sentences, extracting phrases might remove critical context.

4.2 Manual Evaluation

We also applied the three promising pattern matching based methods on articles about COVID-19 from Wikipedia. [2] We evaluated the top 50 outputs from each of the three methods (total 150 outputs) using three annotators experienced in this field [3] and found `PM+NLI`, `PM+CP+NLI` and `PM+NP+NLI` to have a precision of 44.7, 76.7 and 80.7 respectively. We observe that for Wikipedia articles which often have long and complex sentence structures, `PM+NLI` method often gives non-precise extractions while both `PM+CP+NLI` and `PM+NP+NLI` methods have a high precision.

[2] All outputs from the three methods along with human judgments can be found in the supplementary material.

[3] Overall, we observed 82.2% agreement between annotators with Fleiss's Kappa (Fleiss 1971) of 0.6.

DS	Input	Method	P	R	F
BECauSE	Context only	PM	26.3	36.5	30.6
	Context only	PM + NLI	41.6	27.7	33.3
	Context only	PM + CP + NLI	34.0	23.2	27.6
	Context only	PM + NP + NLI	7.0	3.5	4.6
	Context + Cause + Effect	NLI	74	81.0	77.3
SemEval	Context only	PM	36.1	66.2	46.7
	Context only	PM + NLI	45.9	57.7	51.1
	Context only	PM + CP + NLI	43.3	44.2	43.7
	Context only	PM + NP + NLI	26.6	14.1	18.4
	Context + Cause + Effect	NLI	81.9	87.6	84.7

Table 2: The performance of our methods on two datasets. P, R and F refer to the Precision, Recall and F-score of the different methods. The standard deviation across 5 random runs for all the methods is smaller than 0.6

5 Future Work

In the future we would like to explicitly handle cases (1) in which a cause *prevents* the effect from occurring and (2) where multiple causes may lead to multiple effects. Finally, we are planning to explore the application of our framework in decision support and event forecasting. All our datasets and experimental results will be made publicly available.

References

Dunietz, J.; Levin, L.; and Carbonell, J. 2017. The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, 95–104. Valencia, Spain: Association for Computational Linguistics. doi:10.18653/v1/W17-0812. URL https://www.aclweb.org/anthology/W17-0812.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5): 378.

Handler, A.; Denny, M.; Wallach, H.; and O'Connor, B. 2016. Bag of What? Simple Noun Phrase Extraction for Text Analysis. In *Proceedings of the First Workshop on NLP and Computational Social Science*, 114–124. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/W16-5615. URL https://www.aclweb.org/anthology/W16-5615.

Hassanzadeh, O.; Bhattacharjya, D.; Feblowitz, M.; Srinivas, K.; Perrone, M.; Sohrabi, S.; and Katz, M. 2019. Answering Binary Causal Questions Through Large-Scale Text Mining: An Evaluation Using Cause-Effect Pairs from Human Experts. In Kraus, S., ed., *IJCAI*, 5003–5009.

Hassanzadeh, O.; Bhattacharjya, D.; Feblowitz, M.; Srinivas, K.; Perrone, M.; Sohrabi, S.; and Katz, M. 2020. Causal Knowledge Extraction through Large-Scale Text Mining. In *AAAI*, 13610–13611.

Radinsky, K.; Davidovich, S.; and Markovitch, S. 2012. Learning causality for news events prediction. In *WWW*. doi:10.1145/2187836.2187958. URL https://doi.org/10.1145/2187836.2187958.

Sharp, R.; Surdeanu, M.; Jansen, P.; Clark, P.; and Hammond, M. 2016. Creating Causal Embeddings for Question Answering with Minimal Supervision. In *EMNLP*. URL http://aclweb.org/anthology/D/D16/D16-1014.pdf.