Causality Mining in Fiction

Margaret Meehan, Dane Malenfant and Andrew Piper

[1]McGill University, 688 Sherbrooke St, H3A3R1 Montreal, Canada

Abstract

Causality mining has emerged as an important dimension of NLP research, especially with respect to narrative understanding and story generation. Past research in poetics and reading comprehension has identified the expression of causal relations as one of the primary functions of narrative communication. Nevertheless, causality detection remains a formidable challenge for NLP. A recent survey by Ali et al. highlights the many challenges of this interdisciplinary research across multiple domains [1]. One area that remains under-researched is that of cross-domain consistency, whether causal relations and their detection behave in similar ways across different kinds of texts. As Bamman et al. have established, many NLP systems lose considerable performance when applied on fictional texts [2]. In this paper, we thus establish the relative performance of baseline causal mining models on examples drawn from fictional narratives and compare them with standard NLP benchmarks drawn from SemEval data. With a new labeled dataset that we introduce, we train models to detect causality within and between sentences, and uncover linguistic features that indicate a causal relationship between phrases. This research surfaces a range of questions that will help advance the further study of understanding causality in narrative.

Keywords

causality, narratology, digital humanities, natural language processing, fiction, BERT

1. Introduction

Understanding narrative as a form of causal reasoning first emerged in the work of French structuralists [3, 4]. For Tzvetan Todorov, the causal association within or between statements was one of the three fundamental dimensions of narrative. "Most works of fiction of the past are organized according to an order that we may qualify as both temporal and logical;...the logical relation we habitually think of as implication, or as we ordinarily say, causality" (p.41) [3]. Subsequent research has highlighted the significance of causality for narrative comprehension [5]. Causality is not only a core feature that aids comprehension for developing readers [6]; readers of all ages recall information more easily when it is encoded within causally linked phrases [7]. Causal statements have been shown to be powerful predictors of reader comprehension [8] and narrative persuasion [9]. As Graesser writes, summarizing much of this research, "Comprehension is driven by why-questions to a much greater extent than other types of questions (when, where, how, what-happens-next)" [10].

Text-based causality has been researched in non-literary contexts as well, for instance to

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'22 Workshop, Stavanger (Norway), 10-April-2022

margaret.meehan@mail.mcgill.ca (M. Meehan); dane.malenfant@mail.mcgill.ca (D. Malenfant); andrew.piper@mcgill.ca (A. Piper)

0000-0001-9663-5999 (A. Piper)

explore clinical narratives [11] or in social science to analyze policy effects [12]. Within the field of NLP, researchers have studied how causal inference can improve fairness in AI models [13] and shown how causal relations are important components of event scripts, which are themselves elements in the larger task of plot detection [14, 15, 16]

Causality is first and foremost a way of understanding event relationships [14]. Two or more events can be said to be causally related when a temporally prior action or state causes a subsequent action or state. For van den Broek [17], one of the pioneers of causality research, causality consists of four interlocking criteria:

- `temporal priority`: the causal event must be temporally prior to the caused event;
- `operativity`: the causal event must be active when the caused event occurs (i.e. is not subsumed by events when the caused event occurs);
- `necessity`: the caused event would not have occurred without the causal event;
- `sufficiency`: if given the circumstances of the story, the causal event occurs then the caused event occurs;

While prior work has focused on the task of causality mining, one area that remains understudied is the behavior of models across different text domains. As the work of Bamman et al. has shown [2], the performance of many NLP systems degrades considerably when applied to fictional texts (in some cases by twenty points or more). Before proceeding with the more general task of causality detection, we need to better understand its behavior in fictional compared to non-fictional texts.

In this paper, we introduce a pilot dataset of manually annotated causally-related events in sentences drawn from popular contemporary fictional narratives belonging to seven different genres, available in a GitHub repository [1] We also develop computational models to detect causal relations and compare the relative performance of our models between the literary fiction dataset and a set of non-fiction sentences drawn from the SemEval 2010 task-8 dataset [18]. First, we create a model to detect relations between events by utilizing features derived from semantic, syntactic, and dependency-related aspects of the text. Second, we develop a deep-learning model to detect the presence of a causal relation, using pre-trained BERT layers. We compare our models to state-of-the art models and explore implications for further study of causality in narrative texts.

2. Data

2.1. Annotated Data

We provide a set of 548 positively annotated causal event pairs drawn from 141 literary passages along with an equal number of non-causally related pairs from the same passages. Passages from which the event pairs are drawn consist of three consecutive sentences randomly selected from a larger set of over 1,200 literary works published in the past 15 years comprised of seven different genres [19].

[1]The datasets and annotation handbook are available in a GitHub repository.

Table 1
Literary Data Causal Events

Causal Event	Caused Event	Strength Score
she'd fallen asleep against the window	her forehead was cold	3
my oldest and best friend has a glamorous PR job and is freshly engaged	the situation is all the more dismal	2.75
with both feet safely on the ground	Vonetta became her old self	2.25

In order to annotate causal event pairs, over the course of several weeks we trained a team of three undergraduate annotators all with majors in the humanities. In addition to being provided van den Broek's schema cited above, annotators were instructed to maximally annotate an event such that all consecutive words associated with an event were captured (see Table 1 for examples). We define an event for our purposes as any action undertaken by an agent that actually occurs (or does not occur) in the diegetic universe of the narrative. Negative events (things that don't happen) can be both the causal or caused event in our framework, but hypothetical events cannot ("I would go to the market, but..."). We also condition our event pairs on three-sentence excerpts so we only observe event pairs that are within a maximum of two sentences.

Annotators were also asked to provide a "strength" score in accordance with van den Broek's theory that causality is not a purely binary phenomena but one of degree. For example, some events may be more or less sufficient or necessary for a caused event to occur (while temporal priority and operativity are considered to be binary). We therefore created a 3-point ordinal scale to capture annotator confidence in the sufficiency and necessity of the causal event. Figure 1 provides a distribution of scores indicating that in over half of cases causal events were identified as being "strongly" causal. This suggests that the use of an ordinal scale may be extraneous and is worth further study. Overall, we observe moderate agreement between annotators, with Fleiss's Kappa $\kappa = 0.49484$, in line with the inter-agreement score found in the semantic link annotation framework introduced by Mostafazadeh et al. [14].

2.2. SemEval Dataset

We utilize the SemEval 2010 Task-8 dataset, commonly used by Natural Language Processing researchers to train state-of-the-art causality mining algorithms. The SemEval data contains 1003 causal-effect relations. The sentences are manually collected through pattern-based Web search, examples shown in Table 2. The SemEval researchers' definition of cause-effect aligns with ours, although they allow only a specific set of syntactic patterns in the causal relation. For more details see their public annotation guidelines [18].

We combine the cause-effect samples with 1003 rows that do not contain a causal event, in order to capture non-causal relationships. While SemEval annotates events at the word-level, our literary data annotates at the phrase level. In order to compare these two datasets, we computationally capture the phrase around the annotated event. After pre-processing, each row in our dataset contains the causal phrase, the caused phrase, and the sentence.

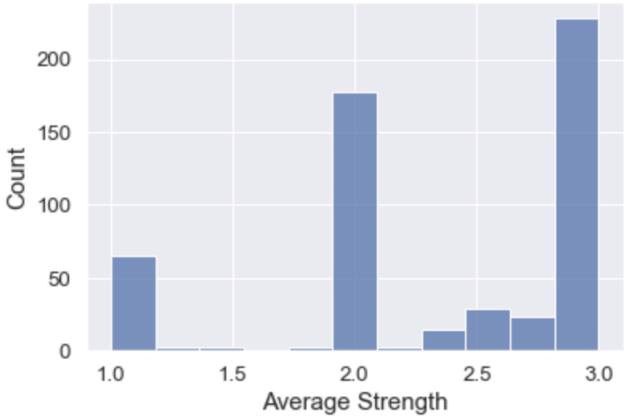

Figure 1: Distribution of literary data causal strength score for all positively annotated events

Table 2
SemEval 2010 Task-8 Annotation Examples

Sentence	Causal Event	Caused Event
Those were cancers caused by radiation exposures.	cancers	exposures
He had chest pains and headaches from mold in the bedrooms.	headaches	mold

3. Models

The goal of causality mining for narrative understanding can assume two distinguishable and important forms. The first is the detection of causal relatedness. The second is the identification of causal pairs. Causal relatedness refers to the identification of the presence of causality communicated in language without identifying the exact two causal units.

While less precise than causal pair detection, the identification of causal relatedness is a valuable form of knowledge with respect to narrative texts. It allows researchers to identify texts, text types (genres), or textual communities (online platforms) where causal argumentation exists in more explicit and detectable ways. For example, we assume that children's literature makes causal relatedness more readily apparent to readers through the explicit use of causally related language. As texts rise in complexity, we assume such causal language will decline, allowing readers more interpretive freedom as to the causes of narrative progression.

The second task, identifying causal pairs, is the more traditional NLP task. This task is important because it can help researchers identify the particular causal relations between agents and events in any narrative environment. For example, we assume that different communities might attribute the causes of Covid to different agents, and the accurate identification of causal pairs could help surface these different community-based narrative understandings.

To address these two tasks we accordingly develop two separate modeling approaches for causal pair detection and causal relation identification.

3.1. Causal Pair Detection

To detect causal pairs, we build a feature set of syntactic, semantic, and dependency-driven representations of the causal phrases in each dataset. We utilize all phrasal pairs from both intra-sentence and inter-sentence causal events. We then experiment with widely used algorithms to train a model on the feature set, and interpret the feature weights of the best model, assessed by the F1 score.

To construct our experimental feature space, we aggregate our features into three general categories:

- `semantic features`: To capture similarity in meaning between two phrases, we calculate the ratio of similar words using a Word2Vec model.
- `syntactic features`: To capture syntactical structure of the two phrases, we maintained a count for each part-of-speech present in the causal and caused phrase. We also engineer a co-reference count between the two phrases, and overlapping entity counts.
- `dependency relationships`: For sentence-level phrase relationships, we count dependency relations between two phrases.

We utilize Python libraries *spaCy* and *neuralcoref* for feature creation [2]. While similar causal research emphasizes the value of temporal relations in detecting causality [14], we lean only on the linguistic qualities within each phrase, or causal unit, rather than such contextual information.

We experiment with models in the widely used *sklearn* library: Random Forests, Support Vector Machines, and Logistic Regression [3]. We utilize grid search and cross validation in order to find the optimized parameter set and feature space, resulting in a trained Random Forest model.

3.2. Causal Sentence Detection

In the 2021 paper "Causality Mining in Natural Languages Using Machine and Deep Learning Techniques: A Survey" Ali et al. [1] review the state of research on causal relationships within computational linguistics. While they review NLP's advancements in causal mining using both traditional and deep-learning techniques, they trained on only non-literary texts. Because we aim to uncover how these types of deep-learning models perform on fictive literary data, we trained a single layer on top of a BERT model to detect intra-sentence causality, i.e. the presence of expressed causality within a given sentence. We then trained the same model architecture on SemEval data to compare the performance. Given a sentence, the model returns the probability of the sentence containing a causal event or not and selects the highest probability.

In our exploration of sentence level causal detection, we first attempted to apply Girju's causal structuring to our dataset [20]. These failed to label the literary text due to the necessary semantic constraints of Girju's pattern matching. These constraints follow a strict causal noun-verb-effect noun structure and were based on Jaegwon Kim's ontological framework of entities

[2] Refer to the spaCy and neuralcoref libraries.
[3] Refer to the sklearn library.

Table 3
Causal Detection Model Parameters

Optimizer	Learning Rate	Epsilon	Dropout Probability	Train Batch	Max Word Length	Epochs
Adam	5e-5	1e-8	0.1	32	124	4

and causal relations [21]. Because these lexico-syntactic patterns did not work well on literary data, we opted for a more robust technique in deep-learning.

BERT is a multilayered bidirectional Transformer based encoder [22]. It is pre-trained to represent sequences of characters or symbols in left and right contexts. BERT is then fine-tuned on an additional layer to model downstream tasks [23]. It is widely used for various NLP tasks and has been employed to detect cause-effect relationships with Semeval 2010 Data [24]. However, this architecture has not been trained on literary data nor has narrative causality been used to aid machine comprehension.

Our BERT architecture and hyperparamters are based on the findings from [23] and incorporates the methodology from the base C-BERT model developed from Vivek et al [24]. We implemented our model with the *transformers* and *pytorch* Python libraries [4]. The data was split 60-40 into a training and test set for the holdout method. The model was then fed literary sentences as sequences of tokens, with a binary indicator of the presence of causality representing non zero strength score. The output contexts vectors were then passed through our Softmax activation layer. Our model used backpropagation with Adam on a learning rate of 5e-5 for the binary loss function on batch sizes of 32 over 4 epochs, see Table 3. This procedure was then repeated with SemEval data.

4. Results

For both sub-problems, the literary datasets achieve similar scores to the SemEval dataset (-5.4 points for causal mining), indicating that existing approaches to causal detection in NLP apply reasonably well to fiction data (Table 4). For both datasets, we achieve the best results when we filter out data that has a low (below 2.0) causal strength score as provided by the annotators.

For sentence level causal detection, on both datasets, the models perform very well: the model trained on SemEval 2010 data has an F1 score of 98.1% and the model trained on the literary data has an F1 Score of 92.7%. This is a surprisingly smaller drop in performance than initially anticipated, given prior work [2]. The application of Girju's patterns on this literary data did not generate meaningful results, but the patterns have previously achieved an F1 score of 82% on SemEval data [24]. Previous entity recognition with machine learning models has also achieved less than 70% F1 score with literary data [2]. While the size of the dataset might have led BERT to overfit, we found consistency across trainings with randomly selected batches of the training set. This indicates that BERT is successful in capturing the complexity of causality in the literary data.

[4]Refer to the transformers and pytorch libraries.

Table 4
Model Performance

Model	Data Type	F1 Score
Causal Sentence (BERT)	Literary	92.7%
Causal Sentence (BERT)	SemEval	98.1%
Causal Pair (RF)	Literary	80.2%
Causal Pair (RF)	SemEval	75.1%

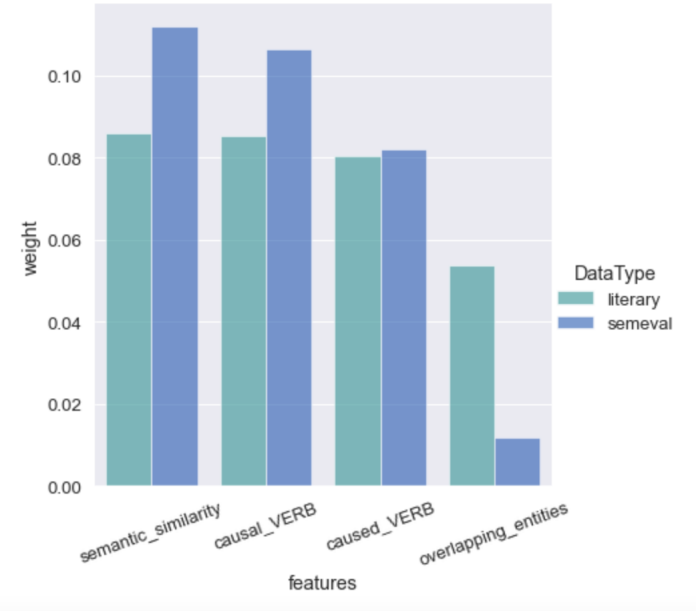

Figure 2: Feature weights for the Random Forest models trained on the annotated literary data and the SemEval data

For causal pair detection using Random Forest models, performance of the linguistic-feature-based models is also shown in Table 4. Rather than prioritizing model performance, we are interested in interpreting the feature weights. In the literary dataset and the SemEval dataset, we find that the number of semantically similar words is the greatest indicator of causality, see Figure 2. This follows our intuition that if an event causes an effect, there will be semantically similar indicator words across the phrases.

As mentioned, applying Girju's pattern matching [20] did not work for the literary dataset, but we found that the presence of certain parts of speech, especially verbs within both the caused and the causal phrase, helped to indicate causality. Certain features that we expected to hold higher weight, like co-reference count, did not, though this might be a bias in the chosen datasets.

5. Discussion

Our models, along with recent research, suggest that causality detection on labeled data is beginning to achieve reasonably high levels of accuracy. When it comes to predicting the presence of causal expression within a given sentence, current deep-learning models are extremely accurate, with only a small decrease in performance for literary data.

Increasing model performance and model confidence however is limited by the relative lack of training data and the complexity of annotating causality [14]. In our literary dataset, we find examples where multiple events cause a subsequent event, or examples in which "necessity" of the relation could be disputed in varying contexts. In our data, we extracted three sentences from longer text passages that when taken out of context might impact reading comprehension. We also only test on limited sentence windows, ignoring causal relatedness across large text spans, which is arguably an important issue for long literary narrative.

In future research we would like to apply a tighter annotation framework to the labeling of events following the schema of Mostafazadeh et al [14], which focuses on single word-phrases that may be hierarchically nested. This would allow for more systematic comparison across multiple domains. Further exploration of inter-annotator agreement and the validity of an ordinal scale for capturing causal relations is also warranted.

6. Conclusion

Despite the inherent challenges in tagging the data, our study finds that it is possible to detect the presence of causal relations in literary sentences utilizing pre-trained BERT layers, with similar success to the SemEval data. We achieve less of a drop in the F1 Score than suggested by prior work with other semantic features [2]. We also show that in detecting the causal phrase, consistent feature weights are used across the two datasets. This indicates that we should be able to use preexisting architectures developed by the NLP community to detect causal event relationships between phrases in literary data. In further research, we aim to apply these architectures for the detection of causal pairs in literary data for both intra-sentence and inter-sentence phrase pairs. This would allow us to perform causal network analysis in a text. Across larger extracts of text, we would expect temporal understanding, using a schema like TimeML, to be more essential [25].

This paper has established the relative performance of baseline models on our literary fiction data versus the commonly-used SemEval dataset. We confirm that causality is a detectable feature in fiction data, and that the models weigh the linguistic features similarly. Even with a small set of hand-annotated training data, we are able to compare our model to state-of-the-art results from the field. This will allow researchers to utilize the model architectures developed in the Natural Language Processing field to study causality in literary texts. How causal relations are encoded differently across different texts, text types, or textual communities remains an open question, but one that might be approachable given our findings.

References

[1] W. Ali, W. Zuo, R. Ali, X. Zuo, G. Rahman, Causality mining in natural languages using machine and deep learning techniques: A survey, Applied Sciences 11 (2021). URL: https://www.mdpi.com/2076-3417/11/21/10064. doi:10.3390/app112110064.

[2] D. Bamman, S. Popat, S. Shen, An annotated dataset of literary entities, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2138–2144. URL: https://aclanthology.org/N19-1220. doi:10.18653/v1/N19-1220.

[3] T. Todorov, Introduction to Poetics, Theory and history of literature, University of Minnesota Press, 1981. URL: https://books.google.com/books?id=IKWaFQZ5xf0C.

[4] R. Barthes, L. Duisit, An introduction to the structural analysis of narrative, New Literary History 6 (1966) 237.

[5] T. Trabasso, P. van den Broek, Causal thinking and the representation of narrative events, Journal of Memory and Language 24 (1985) 612–630. URL: https://www.sciencedirect.com/science/article/pii/0749596X8590049X. doi:https://doi.org/10.1016/0749-596X(85)90049-X.

[6] K. Beker, D. D. Jolles, P. van den Broek, Chapter 2. meaningful learning from texts: The construction of knowledge representations, 2017.

[7] T. Trabasso, L. L. Sperry, Causal relatedness and importance of story events, Journal of Memory and Language 24 (1985) 595–611. URL: https://www.sciencedirect.com/science/article/pii/0749596X85900488. doi:https://doi.org/10.1016/0749-596X(85)90048-8.

[8] P. van den Broek, E. Lorch, R. Thurlow, Children's and adults' memory for television stories: The role of causal factors, story-grammar categories, and hierarchical level, Child development 67 (1997) 3010–28. doi:10.2307/1131764.

[9] M. F. Dahlstrom, The role of causality in information acceptance in narratives: An example from science communication, Communication Research 37 (2010) 857–875. URL: https://doi.org/10.1177/0093650210362683. doi:10.1177/0093650210362683. arXiv:https://doi.org/10.1177/0093650210362683.

[10] A. C. Graesser, B. A. Olde, B. Klettke, How does the mind construct and represent stories, 2002.

[11] I. Spasic, G. Nenadic, Clinical text data in machine learning: Systematic review, JMIR Med Inform 8 (2020) e17984. URL: http://medinform.jmir.org/2020/3/e17984/. doi:10.2196/17984.

[12] N. Egami, C. J. Fong, J. Grimmer, M. E. Roberts, B. M. Stewart, How to make causal inferences using texts, 2018. arXiv:1802.02163.

[13] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, Avoiding discrimination through causal reasoning, 2018. arXiv:1706.02744.

[14] N. Mostafazadeh, A. Grealish, N. Chambers, J. F. Allen, L. Vanderwende, Caters: Causal and temporal relation scheme for semantic annotation of event structures, in: EVENTS@HLT-NAACL, 2016.

[15] N. Chambers, D. Jurafsky, Unsupervised learning of narrative schemas and their participants, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL

and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, ACL '09, Association for Computational Linguistics, USA, 2009, p. 602–610.

[16] N. Chambers, D. Jurafsky, Unsupervised learning of narrative event chains, in: Proceedings of ACL-08: HLT, Association for Computational Linguistics, Columbus, Ohio, 2008, pp. 789–797. URL: https://aclanthology.org/P08-1090.

[17] P. van den Broek, The causal inference maker: Towards a process model of inference generation in text comprehension, Comprehension Processes in Reading (1990).

[18] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, S. Szpakowicz, Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals, CoRR abs/1911.10422 (2019). URL: http://arxiv.org/abs/1911.10422. arXiv:1911.10422.

[19] A. Piper, E. Portelance, How cultural capital works: Prizewinning novels, bestsellers, and the time of reading, Post45 10 (2016).

[20] C. R. Girju, D. Moldovan, Text Mining for Semantic Relations, Ph.D. thesis, 2002. AAI3049820.

[21] J. Kim, Causes and events: Mackie on causation, Journal of Philosophy 68 (1971) 426–441. doi:10.2307/2025175.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[24] V. Khetan, R. R. Ramnani, M. Anand, S. Sengupta, A. E. Fano, Causal-bert : Language models for causality detection between events expressed in text, CoRR abs/2012.05453 (2020). URL: https://arxiv.org/abs/2012.05453. arXiv:2012.05453.

[25] J. Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz, Timeml: A specification language for temporal and event expressions (2003).