Event Causality Identification - Shared Task 3, CASE 2023

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

Hansi Hettiarachchi
School of Computing and Digital Technology, Birmingham City University, United Kingdom
hansi.hettiarachchi@mail.bcu.ac.uk

Ali Hürriyetoğlu
KNAW Humanities Cluster DHLab, The Netherlands
ali.hurriyetoglu@dh.huc.knaw.nl

Nelleke Oostdijk
Centre for Language Studies, Radboud University, The Netherlands
nelleke.oostdijk@ru.nl

Onur Uca
Department of Sociology, Mersin University, Turkey
onuruca@mersin.edu.tr

Surendrabikram Thapa
Department of Computer Science, Virginia Tech, United States of America
surendrabikram@vt.edu

Farhana Ferdousi Liza
School of Computing Sciences University of East Anglia, United Kingdom
f.liza@uea.ac.uk

Abstract

The Event Causality Identification Shared Task of CASE 2023 is the second iteration of a shared task centered around the Causal News Corpus. Two subtasks were involved: In Subtask 1, participants were challenged to predict if a sentence contains a causal relation or not. In Subtask 2, participants were challenged to identify the Cause, Effect, and Signal spans given an input causal sentence. For both subtasks, participants uploaded their predictions for a held-out test set, and ranking was done based on binary F1 and macro F1 scores for Subtask 1 and 2, respectively. This paper includes an overview of the work of the ten teams that submitted their results to our competition and the six system description papers that were received. The highest F1 scores achieved for Subtask 1 and 2 were 84.66% and 72.79%, respectively.

Keywords: Causal News Corpus, Causal event classification, Cause-Effect-Signal span detection

1 Introduction

A causal relation represents a semantic relationship between a Cause argument and an Effect argument, where the occurrence of the Cause triggers the occurrence of the Effect (Barik et al., 2016). The extraction of causal information from text holds significant implications for downstream applications in natural language processing (NLP), like for summarization and prediction (Radinsky et al., 2012; Radinsky and Horvitz, 2013; Izumi et al., 2021; Hashimoto et al., 2014), question answering (Dalal et al., 2021; Hassanzadeh et al., 2019; Stasaski et al., 2021), inference and understanding (Jo et al., 2021; Dunietz et al., 2020).

Given the limited availability of data for causal text mining (Asghar, 2016; Xu et al., 2020; Yang et al., 2022; Tan et al., 2022b), in 2022, the Causal News Corpus (CNC) was created (Tan et al., 2022b).[1] We also introduced a shared task to promote modelling for two causal text mining tasks: (1) Causal Event Classification and (2) Cause-Effect-Signal Span Detection (Tan et al., 2022a). This paper describes the second iteration of this shared task. In this iteration, some parts of our data have updated labels and for Subtask 2, much more annotated data is provided.

The remainder of the paper is organized as follows: Section 2 describes the dataset and its annotations. Section 3 formally introduces the two subtasks for the shared task. Section 4 describes the evaluation metrics and competition set-up. Next, Section 5 summarizes the methods used by par-

[1]The CNC was created by a similar group of authors, some of which did not work on this shared task.

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 144–150, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_019

Stat.	Label	Train	Dev	Test	Total
# Sentences	*Causal*	1624	185	173	1982
	Non-causal	1451	155	179	1785
	Total	3075	340	352	3767
Avg. # words	*Causal*	33.44	34.41	35.93	33.75
	Non-causal	26.69	26.85	28.67	26.90
	Total	30.25	30.96	32.24	30.50

Table 1: Subtask 1 Data Summary Statistics.

Statistic	Train	Dev	Test	Total
# Sentences	1624	185	173	1982
# Relations	2257	249	248	2754
Avg. rels/sent	1.39	1.35	1.43	1.39
Avg. # words	33.44	34.41	35.93	33.75
Cause	11.56	12.20	12.96	11.74
Effect	10.71	10.18	11.54	10.74
Signal	1.45	1.53	1.46	1.46
Avg # *Sig.*/rel	0.70	0.64	0.79	0.70
Prop. of rels w/ *Sig.*	0.68	0.63	0.76	0.69

Table 2: Subtask 2 Data Summary Statistics.

ticipants in the competition. Finally, Section 6 concludes this paper.

2 Dataset

Our shared task uses Version 2 (V2) of the Causal News Corpus (Tan et al., 2022b), which is based on the corpora released in the scope of Hürriyetoğlu et al. (2021).[2] V2 incorporates additional span annotations for Subtask 2. As compared to the previous version of 160 sentences and 183 relations, the current version contains 1981 sentences and 2754 causal relations. Annotations were also revised for some examples across both Subtasks. The summary statistics for Subtask 1 and 2 are available in Tables 1 and 2 respectively.

3 Shared Task Description

The task is comprised of two subtasks related to Event Causality Identification: (1) Causal Event Classification and (2) Cause-Effect-Signal Span Detection. The objective of each subtask is described below in Sections 3.1 and 3.2. The 2023 edition is the second iteration of this shared task which was first introduced in 2022 (Tan et al., 2022a). The shared task is re-launched to work on the larger and revised CNC-V2 discussed in the earlier Section. Additionally, for Subtask 2, the traditional evaluation metrics (P, R and F1) were updated to use fairer evaluation calculations, discussed in Section 4.1.

3.1 Subtask 1: Causal Event Classification

The aim of this task is to classify whether an event sentence contains any cause-effect meaning. Systems had to predict *Causal* or *Non-causal* labels per test sentence. An event sentence was defined to be *Causal* if it contains at least one causal relation.

3.2 Subtask 2: Cause-Effect-Signal Span Detection

The objective of this task is to detect the consecutive spans relevant to a *Causal* relation. There are three types of spans involved in a *Causal* relation: The *Cause* span refers to words that describe the event that triggers another *Effect* event. The *Effect* span refers to words that describe the resulting event arising from a *Cause* event. *Signals* are optionally present, and are words that explicitly indicate a *Causal* relation is present. In our dataset, multiple *Causal* relations can exist in a sentence, and participants have to identify all of them.

4 Evaluation & Competition

4.1 Evaluation Metrics

Evaluation metrics were the same as the shared task launched last year (Tan et al., 2022a). For Subtask 1, Precision (P), Recall (R), F1, Accuracy (Acc) and Matthews Correlation Coefficient (MCC) metrics were used. For Subtask 2, Macro P, R and F1 were used. Evaluation was conducted at the relation level. In other words, examples with multiple causal relations were unpacked and each relation contributed equally to the final score. We designed an evaluation algorithm that allows participants to submit multiple Cause-Effect-Signal span predictions per input sequence in any order. One change from the previous years' evaluation is that we use the FairEval implementation[3] of `seqeval` (Nakayama, 2018; Ramshaw and Marcus, 1995) in Subtask 2 to prevents double penalties of close-to-correct predictions (Ortmann, 2022).

4.2 Baseline

For Subtask 1, we replicate last year's BERT benchmark (Tan et al., 2022b,a). The model fine-tunes the pre-trained (PTM) Bidirectional Encoder Representations from Transformers (BERT) model

[2] https://github.com/tanfiona/CausalNewsCorpus

[3] https://huggingface.co/spaces/hpi-dhc/FairEval

Rank	Team Name	Codalab Username	R	P	F1	Acc	MCC
1	-	DeepBlueAI	86.13	83.24	**84.66**	**84.66**	**69.37**
2	InterosML (Patel, 2023)	rpatel12	87.28	81.62	84.36	84.09	68.37
3	BoschAI (Schrader et al., 2023)	timos	87.86	80.00	83.75	83.24	66.83
4	CSECU-DSG (Hossain et al., 2023)	csecudsg	85.55	80.00	82.68	82.39	64.95
5	-	elhammohammadi	**89.60**	76.35	82.45	81.25	63.52
6	BERT Baseline	tanfiona	89.02	75.86	81.91	80.68	62.37
7	Anonymous	sgopala4	86.13	78.01	81.87	81.25	62.88
8	MLModeler5 (Bhatia et al., 2023)	nitanshjain	87.28	65.37	74.75	71.02	44.83
9	VISU	kunwarv4	52.60	**85.85**	65.23	72.44	48.19
10	-	pakapro	47.40	44.09	45.68	44.60	-10.72

Table 3: Subtask 1 Leaderboard. Ranked by Binary F1. All scores are reported in percentages (%). Highest score per column is in bold.

Rank	Team Name	Codalab Username	Overall			Cause (n=119)			Effect (n=119)			Signal (n=98)		
			R	P	F1	R	P	F1	R	P	F1	R	P	F1
1	BoschAI (Schrader et al., 2023)	timos	**63.98**	**84.42**	**72.79**	**59.66**	**85.28**	**70.20**	**62.88**	**82.76**	**71.46**	**70.44**	**85.36**	**77.18**
2	1Cademy Baseline	tanfiona	59.18	60.25	59.71	54.20	60.92	57.36	59.04	65.98	62.32	64.75	54.75	59.33
3	CSECU-DSG (Hossain et al., 2023)	csecudsg	36.12	40.00	37.96	40.00	42.86	41.38	31.44	33.43	32.40	36.72	44.22	40.12
4	-	pakapro	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: Subtask 2 Leaderboard. Ranked by Overall Macro F1. All scores are reported in percentages (%). Highest score per column is in bold.

(Devlin et al., 2019) for sequence classification. After BERT encodes sentences into word embeddings, the hidden state corresponding to the `[CLS]` token is fed through a binary classification head to obtain the predicted logits. We used the `bert-base-cased` pre-trained model.

For Subtask 2, we replicate the top submission from last year's shared task. Team 1Cademy (Chen et al., 2022)[4] framed the challenge as a reading comprehension task that aims to predict the start and end token positions of each Cause, Effect, and Signal span. We used the `albert-xxlarge-v2` (Lan et al., 2019) pre-trained model.

4.3 Competition Set-up

We used the Codalab website to host our competition.[5]

Registration 29 participants requested to participate on the Codalab page. However, we required participants to email us some personal details (Name, Institution and Email) to avoid teams from creating multiple accounts to cheat. Eventually, only 23 participants were successfully registered, out of which, only 10 accounts participated by uploading predictions.

Trial and Test Periods The trial period started on May 01, 2023, where the training and validation data were released. Participants could upload any number of submissions against the validation set, and they could also submit results for the validation set at any point in time. The main purpose of this setting is for participants to familiarise themselves with the Codalab platform.

The test period started on June 15, 2023 and ended on July 7, 2023. Each participant was allowed only 5 submissions to prevent participants from over-fitting to the test set. After the competition ended, an additional scoring page was created,[6] where participants could upload one result a day to generate more scores for their description papers. None of the scores from this additional scoring page were included into the final leaderboard.

For both Subtasks, the performance was ranked by F1 score: the binary F1 score for Subtask 1, and the Macro F1 score for Subtask 2.

5 Participant Systems

5.1 Overview

Nine participants successfully submitted scores to Subtask 1 while only three successfully submitted scores to Subtask 2 during test period. Table 3 and

[4] https://github.com/Gzhang-umich/1CademyTeamOfCASE

[5] The competition page is at https://codalab.lisn.upsaclay.fr/competitions/11784.

[6] The additional scoring page is at https://codalab.lisn.upsaclay.fr/competitions/14265.

Rank	Team Name	Codalab Username	Overall			Cause (n=119)			Effect (n=119)			Signal (n=98)		
			R	P	F1	R	P	F1	R	P	F1	R	P	F1
1	BoschAI (Schrader et al., 2023)	timos	**53.47**	**82.59**	**64.91**	**47.39**	**82.52**	**60.20**	**50.41**	**80.26**	**61.93**	**64.68**	**84.97**	**73.45**
2	1Cademy Baseline	tanfiona	38.68	41.98	40.26	33.64	40.45	36.73	36.04	43.96	39.60	47.00	41.59	44.13
3	CSECU-DSG (Hossain et al., 2023)	csecudsg	21.16	24.80	22.84	24.63	26.46	25.51	14.66	16.97	15.73	23.96	31.51	27.22
4	-	pakapro	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5: Subtask 2 Leaderboard for Examples with Multiple Causal Relations. This leaderboard was not used in the competition ranking but provided here for discussion purposes. All scores are reported in percentages (%). Highest score per column is in bold.

4 reflects the leaderboard for Subtask 1 and 2 respectively for evaluation metrics described earlier in Section 4.1. For Subtask 2, we further provided the performance for each span type (i.e., Cause, Effect and Signal). We also provide a separate leaderboard for examples with multiple causal relations in Table 5.

For Subtask 1, the top performing team was DeepBlueAI, scoring 84.66% for F1. DeepBlueAI also topped the charts for Acc and MCC scores. Team InterosML (Patel, 2023) followed closely after, with an F1 score of 84.36%. Unfortunately, DeepBlueAI did not submit a paper, so we do not know the method they used. InterosML's (Patel, 2023) employed a two-phased approach to fine-tune the model first using RoBERTa embeddings and with contrastive loss.

For Subtask 2, the top performing team was BoschAI (Schrader et al., 2023) with an F1 score of 72.79%, far higher than the 1Cademy baseline that we provided. A key modelling decision that they had was to stack multiple token labels into one target label, thereby allowing their model to detect multiple causal relations per sequence. This key feature sets them ahead of the model design of the 1Cademy baseline. This can be observed by the large improvements in overall F1 score of 24.65% for examples with multiple causal relations in Table 5 (40.26% vs 64.91.%).

All participants used pre-trained models in their frameworks. For Subtask 1, although multiple teams described a similar sequence classification framework using BERT and RoBERTa, different F1 scores were reported. This suggests the importance of carefully designing and implementing suitable hyperparamters in training a model.

5.2 Methods

We summarize the systems of the six teams that submitted description papers below, sorted according to their leaderboard ranking. Only four papers were accepted to be included in the proceedings of the CASE workshop.

5.2.1 Subtask 1

InterosML (Patel, 2023)'s methodology involved two phases: (1) pre-training a baseline RoBERTa model with supervised contrastive loss (SuperCon), and (2) Fine-tuning the pre-trained model on Subtask 1 itself. For Phase 1, the positive instances refer to sequences containing causal relations, while negative instances refer to sequences without causal relations. The authors demonstrate the usefulness of using contrastive loss, achieving high F1 score of 84.36%, clinching 2nd place, and only slightly below the first place's score. In their paper, they present T-SNE visualizations to investigate the effectiveness of their model on the classification task.

BoschAI (Schrader et al., 2023) used a sequence classification framework that outputs a prediction based on the [CLS] embedding. They experimented with two pre-trained models, BERT-large and RoBERTa-large. A weighted cross-entropy loss was applied to up-weight positive samples.

CSECU-DSG (Hossain et al., 2023) used two transformer models, DeBERTa and RoBERTa to extract contextualized embeddings, which are then combined through a linear feed-forward layer to estimate the probability score of each class. A weighted average of the scores from the two modules is used to obtain the final probability of the scores for each label.

Anonymous they experimented with two models: (1) BERT-base sequence classifier and (2) few-shot prompting of GPT-4 using 0, 2, 4, 6, 14 prompts. In their experiments, they showed that a fine-tuned BERT classifier obtains an F1 score of 81.8%, exceeding the best score possible with GPT-4 of 70.7%. They also did not find a correlation between increasing the number of prompts shown to GPT-4 with any improvements in F1.

MLModeler5 (Bhatia et al., 2023) used a RoBERTa sequence classification model to clas-

sify input sequences with a binary label indicating if causal relations exists in the sequence or not. Their main contribution is the exploration of four datasets, created by processing the original data with four different heuristics-based method. According to their experiments, their model performed best when trained on a dataset that had stop words removed and abbreviations were replaced in the input sequences.

VISU used multiple embedding methods (static, stacked, and contextualized) for this task. For non-contextualized embeddings, a BiLSTM was applied onto various embeddings from GloVe, fastText or frozen-BERT. For contextualized embeddings, a linear layer was applied onto various embeddings like ERT-base, BART, DistilBERT or RoBERTa. In their experiments, they demonstrate that contextualized embeddings obtain the highest F1 scores, the best being RoBERTa which scored an F1 of 65.23%.

5.2.2 Subtask 2

BoschAI (Schrader et al., 2023) approached the task as a sequence tagging task using the BILOU (Alex et al., 2007) labeling scheme. This scheme extends the BIO scheme by adding markers for the end of a multi-token sequence (L) and a single-token entity (U). They experimented with two pretrained models, BERT-large and RoBERTa-large, that generate embeddings fed to a linear layer to obtain logits per token, then the logits were parsed through a CRF output layer to compute the most likely consistent tag sequence. However, this approach can only predict a single output sequence per sample, which is not suitable for sentences with multiple causal chains. To address this, the BILOU labels are stacked using a pipe (|) operator similar to Straková et al. (2019), allowing the model to consider multiple causal relations within a single instance. Three layers are used to keep the label space manageable. Stacked labels occurring in the training and validation data are added, resulting in approximately 300 three-layer BILOU labels. During evaluation, these stacked labels are split into three distinct layers, allowing the model to predict up to three different causal relations per sentence. Data augmentation was also used to increase the number of training samples. This approach was able to rank first in the subtask with an F1-score of 72.79%.

CSECU-DSG (Hossain et al., 2023) employed two different transformer models, namely DeBERTa and DistilRoBERTa, independently for capturing cause-effect and signal span features, respectively. Subsequently, they combined both sets of features and fed them into a stacked BiLSTM network to capture long-term relationships among the tokens. After the BiLSTM network, a max-pooling layer and classifier were incorporated to predict token labels. To enhance system performance, the authors introduced a contrastive loss for cause-effect token classification, whereas, for signal token classification, they utilized cross-entropy loss, considering that signal tokens may or may not be present in the text. The R, P, and F1 achieved by the approach were 36.12%, 40.00%, and 37.96% respectively.

6 Conclusion

In conclusion, our shared task investigated two important tasks in causal text mining, namely: (1) Causal Event Classification, and (2) Cause-Effect-Signal Span Detection. Our shared task attracted 23 registered participants and 10 active participants. Based on the six description papers received, some novel methods that exceeded our initial baseline were proposed. The best F1 scores achieved for Subtask 1 and 2 were 84.66% and 72.79% respectively.

References

Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Biological, translational, and clinical language processing*, pages 65–72.

Nabiha Asghar. 2016. Automatic extraction of causal relations from natural language texts: a comprehensive survey. *arXiv preprint arXiv:1605.07895*.

Biswanath Barik, Erwin Marsi, and Pinar Öztürk. 2016. Event causality extraction from natural science literature. *Res. Comput. Sci.*, 117:97–107.

Amrita Bhatia, Ananya Thomas, Nitansh Jain, and Jatin Bedi. 2023. MLModeler5 @ Causal News Corpus 2023: Using roberta for casual event classification. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Varna, Bulgaria (Hybrid). Association for Computational Linguistics.

Xingran Chen, Ge Zhang, Adam Nik, Mingyu Li, and Jie Fu. 2022. 1Cademy @ causal news corpus 2022: Enhance causal span detection via beam-search-based position selector. In *Proceedings of the 5th Workshop on Challenges and Applications of*

Automated Extraction of Socio-political Events from Text (CASE), pages 100–105, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dhairya Dalal, Mihael Arcan, and Paul Buitelaar. 2021. Enhancing multiple-choice question answering with causal knowledge. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 70–80, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997, Baltimore, Maryland. Association for Computational Linguistics.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2019. Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5003–5009. International Joint Conferences on Artificial Intelligence Organization.

MD. Akram Hossain, Abdul Aziz, and Abu Nowshed Chy. 2023. CSECU-DSG @ Causal News Corpus 2023: Leveraging roberta and deberta transformer model with contrastive learning for causal event classification. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Varna, Bulgaria (Hybrid). Association for Computational Linguistics.

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Farhana Ferdousi Liza, Ritesh Kumar, and Shyam Ratan. 2021. Multilingual protest news detection - shared task 1, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 79–91, Online. Association for Computational Linguistics.

Kiyoshi Izumi, Hitomi Sano, and Hiroki Sakaji. 2021. Economic causal-chain search and economic indicator prediction using textual data. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 19–25, Lancaster, United Kingdom. Association for Computational Linguistics.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard H. Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Trans. Assoc. Comput. Linguistics*, 9:721–739.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Katrin Ortmann. 2022. Fine-grained error analysis and fair evaluation of labeled spans. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1400–1407, Marseille, France. European Language Resources Association.

Rajat Patel. 2023. InterosML @ Causal News Corpus 2023: Understanding causal relationships. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Varna, Bulgaria (Hybrid). Association for Computational Linguistics.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 909–918. ACM.

Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 255–264. ACM.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Timo Pierre Schrader, Simon Razniewski, Lukas Lange, and Annemarie Friedrich. 2023. BoschAI @ Causal News Corpus 2023: Robust cause-effect span extraction using multi-layer sequence tagging and data augmentation. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Varna, Bulgaria (Hybrid). Association for Computational Linguistics.

Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A. Hearst. 2021. Automatically generating cause-and-effect questions from passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 158–170, Online. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested ner through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Tommaso Caselli, Onur Uca, Farhana Ferdousi Liza, and Nelleke Oostdijk. 2022a. Event causality identification with causal news corpus - shared task 3, CASE 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 195–208, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022b. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Jinghang Xu, Wanli Zuo, Shining Liang, and Xianglin Zuo. 2020. A review of dataset and labeling methods for causality extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1519–1531, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowl. Inf. Syst.*, 64(5):1161–1186.