Event Causality Identification with Causal News Corpus - Shared Task 3, CASE 2022

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`

Tommaso Caselli
CLCG,
University of Groningen,
The Netherlands
`t.caselli@rug.nl`

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

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

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

Abstract

The Event Causality Identification Shared Task of CASE 2022 involved two subtasks working on the Causal News Corpus. Subtask 1 required participants to predict if a sentence contains a causal relation or not. This is a supervised binary classification task. Subtask 2 required participants to identify the Cause, Effect and Signal spans per causal sentence. This could be seen as a supervised sequence labeling task. 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 summarizes the work of the 17 teams that submitted their results to our competition and 12 system description papers that were received. The best F1 scores achieved for Subtask 1 and 2 were 86.19% and 54.15%, respectively. All the top-performing approaches involved pre-trained language models fine-tuned to the targeted task. We further discuss these approaches and analyze errors across participants' systems in this paper.

1 Introduction

A causal relation represents a semantic relationship between a Cause argument and an Effect argument, in which the occurrence of the Cause leads to the occurrence of the Effect (Barik et al., 2016). Extracting causal information from text has many downstream natural language processing (NLP) applications, 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).

However, data for causal text mining is limited (Asghar, 2016; Xu et al., 2020; Yang et al., 2022; Tan et al., 2021, 2022a). There are also not many benchmarks to allow for fair model comparisons (Asghar, 2016). Therefore, in this paper, we continue our efforts with the creation of the Causal News Corpus (CNC). CNC is a corpus of news articles annotated with causal information suitable for causal text mining. Additionally, we introduce a shared task to promote modelling for two causal text mining tasks: (1) Causal Event Classification and (2) Cause-Effect-Signal Span Detection. Figure 1 provides examples from the CNC in this shared task. To our knowledge, we are the first dedicated causal text mining dataset and benchmark to include signal span detection as an objective.

The rest of the paper is organized as follows: Section 2 presents literature on event causality datasets. Section 3 describes the dataset and annotation of the corpus. Section 4 formally introduces the two subtasks for the shared task. Section 5 describes the evaluation metrics and competition set-up. Subsequently, Section 6 summarizes the methods used by
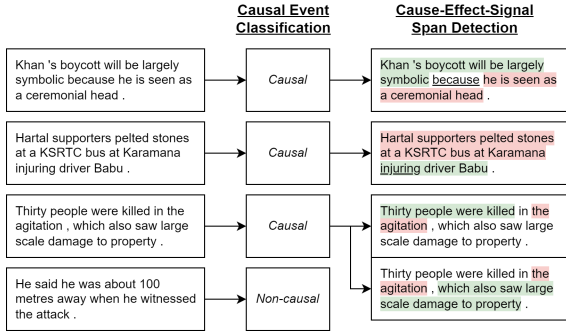
Figure 1: Examples from the CNC for the two subtasks. Cause spans are indicated by Pink, while Effect spans are indicated by Green. Signals, if present, are underlined.

participants during the competition, while Section 7 analyzes the participants' submissions. Finally, Section 8 concludes this paper.

2 Related Work

In many papers about Event Causality Identification (ECI) (Gao et al., 2019; Zuo et al., 2021b; Cao et al., 2021; Zuo et al., 2021a, 2020), the two datasets used for benchmarking are CausalTimeBank (Mirza et al., 2014; Mirza and Tonelli, 2014) and EventStoryLine (Caselli and Vossen, 2017). These datasets are unsuitable for span detection since their arguments are event headwords only.

There are two other efforts that intentionally introduce datasets for benchmarking causal text mining systems. FinCausal (Mariko et al., 2021, 2020) is a recurring shared task held within the FinNLP workshop focusing on financial news. In the first subtask, participants also aim to identify if sentences contain causal relations. In the second subtask, participants to identify the Cause and Effect spans in the causal sentences. UniCausal (Tan et al., 2022b)[1] is an open-source repository for causal text mining that has consolidated six corpora for three causal text mining tasks. The six corpora included in UniCausal are: AltLex (Hidey and McKeown, 2016), BECAUSE 2.0 (Dunietz et al., 2017), CausalTimeBank (Mirza et al., 2014; Mirza and Tonelli, 2014), EventStoryLine V1.0 (Caselli and Vossen, 2017), Penn Discourse Treebank V3.0 (Webber et al., 2019), and SemEval 2010 Task 8 (Hendrickx et al., 2010). The three tasks are: Causal Sentence Classification, Causal Pair Classification and Cause-Effect Span Detection.

Similar to FinCausal and UniCausal, we included a signal span detection objective. Our annotation guidelines differ slightly, in that our arguments must contain events, and the spans are annotated in a manner that is minimally sufficient. In general, we notice that spans from FinCausal are much longer. Spans from UniCausal depend on the original data source.

Additionally, for Cause-Effect Span Detection in FinCausal, their approach to handle multiple causal relations per unique sentence was to include index numbers at the start of each sentence to differentiate the Cause-Effect predictions. This approach is problematic because (1) it leaks information that the sentence contains multiple causal relations to the model, and (2) predictions that are submitted in a different order from the true labels are unnecessarily penalised. Therefore, we differ from FinCausal when evaluating multiple causal relations in span detection since we group relations by its sentence index. This is described further in Section 5.1.

3 Dataset

3.1 Data Collection

Our shared task worked with the Causal News Corpus (CNC) (Tan et al., 2022a)[2], which consists of 869 news documents and 3,559 English sentences, annotated with causal information. CNC builds on the randomly sampled articles (Yörük et al., 2021) from multiple sources and periods featured (Hürriyetoğlu et al., 2021) in a series of workshops directed at mining socio-political events from news articles (Hürriyetoğlu et al., 2020b,a, 2021a,b; Hürriyetoğlu, 2021). CNC follows the train-test split of the original data source, with 3,248 training and 311 test examples. Later, we further split and randomly sampled 10% of the original training set to obtain the development set. Later, Table 3 presents the sentence counts per data split.

3.2 Annotation

3.2.1 Guidelines

For more information on our annotation guidelines, please refer to our annotation manual[3].

Subtask 1 In CNC, sentences were labeled as *Causal* or *Non-causal*, where the presence of causality indicates that "one argument provides the

---

reason, explanation or justification for the situation described by the other" (Webber et al., 2019). Our sentences had to contain at least a pair of events, defined as "things that happen or occur, or states that are valid" (Saurı et al., 2006). These annotations correspond to the target labels for Subtask 1, Causal Event Classification.

Subtask 2 For *Causal* sentences, the words corresponding to the Cause-Effect-Signal spans of a causal relation were also marked. These annotations correspond to the target labels for Subtask 2, Cause-Effect-Signal Span Detection. However, at the current stage of writing, only a small subset of our data contains annotated spans. Span annotations are an on-going effort.

A Cause is a reason, explanation or justification that led to an Effect. We defined a Cause or Effect span as a continuous set of words sufficient for the interpretation of the causal relation meaning. This means that any context modifying or describing the argument relevant to the causal relation was included. Each Cause or Effect span must contain an event, where an event is defined as a situation that 'happen or occur', or predicates that 'describe states or circumstances in which something obtains or holds true' (Pustejovsky et al., 2003).

Signals are words that help to identify the structure of the discourse. In our case, signals highlight the relationship between the Cause and Effect.

3.2.2 Annotation Tool

We used the WebAnno tool (Eckart de Castilho et al., 2016) to conduct our annotation process.

Subtask 1 Annotation at the sequence level was relatively straightforward, where annotators selected "Yes" or "No" labels for each sentence.

Subtask 2 Annotators first marked the Cause span, Effect span, and Signal span. Subsequently, they linked the spans together by pointing Cause to Effect and Signal to Effect. An illustration is provided in Figure 2. Annotations were then downloaded and sent through checking scripts on Python to identify if there were any avoidable human errors. For example, if missing links (E.g. An Effect has no Cause) or invalid links (E.g. An Effect points to Effect) were present, and an error report was then sent to annotators for them to consider correcting their annotations.

	Train	Dev	Test	Total
K-Alpha	34.42	29.77	48.55	34.99

Table 1: Subtask 1 Inter-annotator Agreement Scores. Reported in percentages.

Metric	Span	Train+Dev	Test	Total
Exact Match	Cause	30.57	15.11	23.88
	Effect	36.30	19.86	29.19
	Signal	27.92	29.21	28.48
	Total	7.84	5.81	6.96
One-Side Bound	Cause	57.55	39.86	49.90
	Effect	60.90	45.42	54.21
	Signal	31.93	32.96	32.37
	Total	24.05	22.25	23.27
Token Overlap	Cause	63.65	49.18	57.39
	Effect	64.66	49.88	58.27
	Signal	32.09	33.15	32.55
	Total	26.94	27.78	27.31
K-Alpha	Cause	46.36	42.51	44.32
	Effect	57.18	41.89	49.89
	Signal	29.30	23.42	27.08
	Total	50.90	41.54	46.27

Table 2: Subtask 2 Inter-annotator Agreement Scores. Reported in percentages (%).

3.2.3 Annotation Process & Curation

Five annotators were involved and independently annotated for both subtasks across the span of a few months. For each round of annotations, annotators were presented with a subset of the dataset. After each round, the curator consolidated the final annotations as follows:

Subtask 1 The majority voted label was retained as the final label. Every example in the final corpus was annotated by at least two annotators. The curator has the final vote if there are ties, or if only one annotation is present. Further details are available in the CNC paper (Tan et al., 2022a).

Subtask 2 There was no straightforward way to take a majority label for span annotations. Therefore, our approach was that the curator took into account the spans highlighted by the annotators and decided on the final selection.

After each annotation round, the final span annotations were made available for annotators to review and discuss.

3.2.4 Summary Statistics

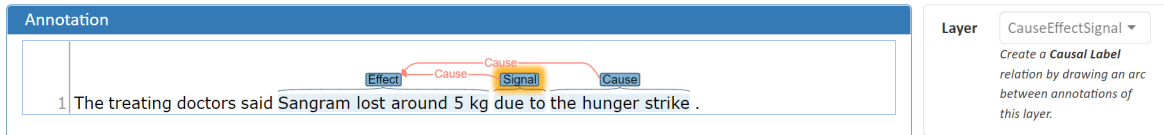**Inter-annotator Agreement** For Subtask 1, scores are reflected in Table 1. Also reported in

Figure 2: Screenshot of the annotation tool used to mark Cause-Effect-Signal spans.

Stat.	Label	Train	Dev	Test	Total
# Sent-ences	*Causal*	1603	178	176	1957
	Non-causal	1322	145	135	1602
	Total	2925	323	311	3559
Avg. # words	*Causal*	35.48	36.86	41.27	36.13
	Non-causal	27.34	27.35	30.25	27.59
	Total	31.80	32.59	36.49	32.28

Table 3: Subtask 1 Data Summary Statistics.

Stat.	Train	Dev	Test	Total
# Sentences	160	15	89	264
# Relations	183	18	119	320
Avg. rels/sent	1.14	1.20	1.34	1.21
Avg. # words	17.21	16.13	28.45	20.94
Cause	6.52	7.28	12.76	8.89
Effect	7.80	6.44	10.20	8.62
Signal	1.55	1.60	1.36	1.47
Avg # signals/rel	0.67	0.56	0.82	0.72
Prop. of rels w/ signals	0.64	0.56	0.76	0.68

Table 4: Subtask 2 Data Summary Statistics.

Tan et al. (2022a), overall, the dataset has a Krippendorff's Alpha (K-Alpha) agreement score of 34.99%.

For Subtask 2, the agreement metrics used were Exact Match (EM), Token Overlap (TO), One-Side Bound (OSB), and K-Alpha. Scores are presented in Table 2. Overall, the dataset had agreement scores of 6.96% EM, 23.27% OSB, 27.31% TO, and 46.27% K-Alpha. Since OSB and TO are relaxed span evaluation metrics (Lee and Sun, 2019), they are naturally much higher than EM, which is a strict metric. How the metrics were calculated is described in the Appendix Section A.1.

Shared Task Data   The summary statistics for Subtask 1 and 2 are available in Tables 3 and 4 respectively.

It is worth noting that for Subtask 2, the test set contained sentences that were much longer than those in the training sets. This is because we were annotating the shorter sentences first based on annotators' feedback that working with shorter sentences at the beginning helps them to familiarise themselves with the annotation rules. Since there were more sentences in the training set, the training set naturally also had more short sentences for us to annotate first. Once we are done with span annotations, the average number of words for Subtask 2 should tally with the causal sentences of Subtask 1, shown earlier in Table 3.

4   Task Description

The shared task is comprised of two subtasks related to Event Causality Identification. The objective of each task is described in detail as follows:

4.1   Subtask 1: Causal Event Classification

The objective 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.

4.2   Subtask 2: Cause-Effect-Signal Span Detection

The objective of this task is the detection of 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.

5   Evaluation & Competition

5.1   Evaluation Metrics

5.1.1   Subtask 1

We evaluated participants' predictions using Accuracy (Acc), Precision (P), Recall (R), F1, and Matthews Correlation Coefficient (MCC) scores.

5.1.2   Subtask 2

Following previous evaluation metrics for Cause-Effect Span Detection (Mariko et al., 2020, 2021) and text chunking (Tjong Kim Sang and Buchholz,

2000), we assessed predictions using Macro P, R and F1 metrics.

Participants uploaded sentences with Cause-Effect-Signal spans marked directly in the text using `ARG0`, `ARG1` and `SIG` start and end boundary markers. We converted these marked sentences into two white-space tokenized sequences, one corresponding to the token labels for Cause and Effect, and another corresponding to the token labels for Signals. We used the token classification evaluation scheme from `seqeval` (Nakayama, 2018; Ramshaw and Marcus, 1995)[4] provided through Huggingface (Wolf et al., 2020)[5].

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.

Handling multiple relations   Since one input sequence can return multiple causal relations, we adjusted the evaluation code to automatically extract the combination that results in the best F1 score. As such, participants could submit multiple Cause-Effect-Signal span predictions per input sequence in any order. An illustration is provided in Figure 5.1.2.

In evaluation, we only compare with the number of causal relations that the true label has. Let the number of predicted relations be n_p, and the number of actual relations be n_a. Our evaluation script does the following:

- If the number of predicted relations exceeds the number of actual relations ($n_p > n_a$), we kept only the first n_a predictions.

- If the number of predicted relations is less than the number of actual relations ($n_p < n_a$), the missing $n_a - n_p, n_a - n_p + 1, ..., n_a$ relation predictions were represented by tokens that all correspond to the Other (`O`) label.

5.2   Baseline

For Subtask 1, we duplicated the BERT (Devlin et al., 2019) and LSTM (Hochreiter and Schmidhuber, 1997) baselines from our previous work (Tan et al., 2022a) that achieved F1 scores of 81.20% and 78.22% respectively.

For Subtask 2, a random baseline[6] was created for reference. This baseline first randomly identifies start positions for Cause and Effect spans, and then identifies end positions for these spans with a linearly increasing probability as we move away from the start location in order to reflect our preference for longer spans. We also randomly predicted words to be signals with a 10% chance. The baseline F1 score was 0.45%.

5.3   Competition Set-up

We used the Codalab website to host our competition.[7]

Registration   37 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. Subsequently, 29 participants were successfully registered, but only 17 accounts participated by uploading predictions.

Trial and Test Periods   The trial period started on April 15, 2022 and the validation labels were released on August 01, 2022. Participants could upload any number of submissions against the validation set, and they could also submit predictions 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 August 01, 2022 and ended on August 31, 2022. 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,[8] where participants could upload one prediction a day to generate more scores for their description papers. Any scores from this additional scoring page is not included into the final leaderboard.

For both subtasks, models were ranked based on F1 performance on the competition test set.

---

[4]https://github.com/chakki-works/seqeval
[5]https://huggingface.co/spaces/evaluate-metric/seqeval

[6]https://github.com/tanfiona/CausalNewsCorpus/blob/master/random_st2.py
[7]The competition page is at https://codalab.lisn.upsaclay.fr/competitions/2299. The additional scoring page is at https://codalab.lisn.upsaclay.fr/competitions/7046.
[8]The additional scoring page is at https://codalab.lisn.upsaclay.fr/competitions/7046.

1. **Example Inputs:** Predicted & Reference marked sentence for one unique sentence.

Predicted	Reference
Violence broke out when <ARG1>DYFI members protested</ARG1> <ARG0><SIG0>against</SIG0> a Youth Congress meeting being held at the junction</ARG0> . **p1**	<ARG1>Violence broke out</ARG1> when <ARG0>DYFI members protested</ARG0> against a Youth Congress meeting being held at the junction . **a1**
Violence <ARG1>broke</ARG1> out when <ARG0>DYFI members protested</ARG0> against a Youth Congress meeting being held at the junction . **p2**	Violence broke out when <ARG1>DYFI members protested</ARG1> <ARG0>against a Youth Congress meeting being held at the junction</ARG0> . **a2**
<ARG1>Violence broke out</ARG1> when <ARG0>DYFI members protested</ARG0> against a Youth Congress meeting being held at the junction . **p3**	

2. **Retain relevant predictions:** We only evaluate against the available number of Reference annotations.

| p1 | p2 | | a1 | a2 |

3. **Evaluate:** Calculate F1 score for Cause, Effect, and Signal. Keep pairing that returns the highest Overall F1 score.

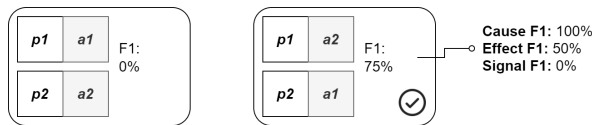| p1	a1	F1: 0%		p1	a2	F1: 75%	Cause F1: 100%
p2	a2			p2	a1	✓	Effect F1: 50%
							Signal F1: 0%

Figure 3: Illustration of how we process multi-relation examples for sequence evaluation.

6 Participant Systems

6.1 Overview

13 participants successfully submitted scores to Subtask 1 while only 4 successfully submitted scores to Subtask 2 during test period. Table 5 and 6 reflects the leaderboard for Subtask 1 and 2 respectively for evaluation metrics described earlier in Section 5.1. For Subtask 2, we further provided the performance for each span type (i.e., Cause, Effect and Signal).

For Subtask 1, the top performing team was CSECU-DSG (Aziz et al., 2022), scoring 86.19% F1. CSECU-DSG also topped the charts for P, Acc, and MCC scores. Team ARGUABLY (Kohli et al., 2022) followed closely after, with 86.10% F1 score and a high recall score of 91.48%. Both methods fine-tuned SOTA pre-trained BERT variants (RoBERTA (Liu et al., 2019) and DeBERTa (He et al., 2021)) to the classification task.

For Subtask 2, the top performing team was 1Cademy (Chen et al., 2022), scoring 54.15% F1. Team IDIAPers (Fajcik et al., 2022) and SPOCK (Saha et al., 2022) followed closely after, with 48.75% and 47.48% F1 scores respectively. Each team approached the span detection task in a different way: 1Cademy treated the task as a reading comprehension challenge and predicted start and end boundaries of the spans. IDIAPers treated the task as a decoding challenge, while SPOCK gener-

ated and classified candidate spans. All participants used pre-trained models in their frameworks.

6.2 Methods

Each teams' systems are summarized below, sorted according to their leaderboard ranking.

6.2.1 Subtask 1

CSECU-DSG (Aziz et al., 2022) proposed a way to unify predictions obtained from two neural network models, by combining the prediction scores generated from each model using a weighted arithmetic mean. The two models used were, Twitter RoBERTa and RoBERTa-base, and each was attached to a linear layer to predict the causal labels. The weights per model were 0.4 and 0.6 respectively, selected through experiments on training data. Their findings on the test set showed that the fused model achieves higher P, R, and F1 score than each model alone, and their approach clinched the top place during the competition.

ARGUABLY (Kohli et al., 2022) proposed using sentence-level data augmentation to fine-tune language models (LMs). They involved contextualised word embeddings of DistilBERT (Sanh et al., 2019) to construct new data. As for the LMs, DeBERTa and dual cross attention RoBERTa models have been experimented with. According to the results, the DeBERTa model fine-tuned on augmented data

Rank	Team Name	Codalab Username	R	P	F1	Acc	MCC
1	CSECU-DSG (Aziz et al., 2022)	csecudsg	88.64	**83.87**	**86.19**	**83.92**	**67.14**
2	ARGUABLY (Kohli et al., 2022)	guneetsk99	**91.48**	81.31	86.10	83.28	66.02
3	LTRC (Adibhatla and Shrivastava, 2022)	hiranmai	88.64	82.11	85.25	82.64	64.51
4	NLP4ITF (Krumbiegel and Decher, 2022)	pogs2022	88.07	82.45	85.16	82.64	64.49
5	IDIAPers (Burdisso et al., 2022)	msingh	87.50	82.80	85.08	82.64	64.49
6	NoisyAnnot (Nguyen and Mitra, 2022)	thearkamitra	88.07	82.01	84.93	82.32	63.83
7	SNU-Causality Lab (Kim et al., 2022)	JuHyeon_Kim	90.34	79.50	84.57	81.35	62.04
8	LXPER AI Research	brucewlee	86.36	82.61	84.44	81.99	63.18
9	1Cademy (Nik et al., 2022)	nika	86.36	81.72	83.98	81.35	61.85
10	-	quynhanh	85.80	79.06	82.29	79.10	57.19
11	BERT Baseline (Tan et al., 2022a)	tanfiona	84.66	78.01	81.20	77.81	54.52
12	GGNN (Trust et al., 2022)	PaulTrust	88.07	74.88	80.94	76.53	52.05
13	LSTM Basline (Tan et al., 2022a)	hansih	84.66	72.68	78.22	73.31	45.15
14	Innovators	lapardnemihk9989	78.98	72.02	75.34	70.74	39.81
15	-	necva	81.25	59.09	68.42	57.56	9.44

Table 5: 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	Acc	R	P	F1	R	P	F1	R	P	F1
1	1Cademy (Chen et al., 2022)	gezhang	53.87	55.09	**54.15**	**43.15**	55.46	57.98	**56.47**	55.46	57.14	**56.13**	50.00	49.09	48.92
2	IDIAPers (Fajcik et al., 2022)	msingh	47.62	51.21	48.75	40.83	45.38	45.38	45.38	42.86	42.86	42.86	**56.12**	**68.44**	**60.01**
3	SPOCK (Saha et al., 2022)	spock	43.75	**57.62**	47.48	36.87	37.82	49.19	41.40	39.50	**59.66**	46.29	56.12	65.39	56.32
4	LTRC (Adibhatla and Shrivastava, 2022)	hiranmai	5.65	2.34	3.23	33.03	2.52	1.10	1.53	13.45	5.51	7.60	0.00	0.00	0.00
5	Random Baseline	tanfiona	0.30	0.89	0.45	21.94	0.84	2.52	1.26	0.00	0.00	0.00	0.00	0.00	0.00

Table 6: Subtask 2 Leaderboard. Ranked by Overall Macro F1. All scores are reported in percentages (%). Highest score per column is in bold.

outperformed the unaugmented DeBERTa model and RoBERTa models.

LTRC (Adibhatla and Shrivastava, 2022) used various transformers-based language models followed by a classification head. The pre-trained models explored by them were: BART-large (Lewis et al., 2020), RoBERTa-base+Linear Layer, RoBERTa-large+Linear Layer, RoBERTa-base+Adapter and RoBERTa-large+Adapter. Their best model slightly beats the baseline scores on the development set.

NLP4ITF (Krumbiegel and Decher, 2022) proposed building a RoBERTa model with linguistic features. They mainly involved named entities (NE) and cause-effect-signal (CES) spans from Subtask 2 to incorporate linguistic features with the input text. Based on their findings, the model trained with the PER (person) NE class with CES, achieved the best results, outperforming the RoBERTa baseline (model trained on data with no linguistic features).

IDIAPers (Burdisso et al., 2022) proposed a prompt-based approach for fine-tuning LMs in which the classification task is modeled as a masked language modeling problem (MLM). This approach allows LMs natively pre-trained on MLM

problems, like RoBERTa, to directly generate textual responses to domain-specific prompts. This approach allow the model to be trained in a few-shot configuration, keeping most of available data for measuring the generalization power the model. The best-performing model was trained with only 256 instances per class and yet was able to obtain the second-best precision and third-best accuracy.

NoisyAnnot (Nguyen and Mitra, 2022) proposed fine-tuning different LMs with customised cross-entropy loss functions that exploit annotation information such as the number of annotators and their agreement. They used several language models including BERT, RoBERTa and XLNET models and showed that the involvement of annotation information improves the model performance.

SNU-Causality Lab (Kim et al., 2022) proposed fine-tuning an ELECTRA model using the CNC dataset and augmented data. They followed two approaches for data augmentation: (1) concatenating SemEval-2010 to CNC and (2) generating new samples using POS tagging. With the POS tagging-based approach they mainly targeted replacing causality irrelevant words with POS tags, to generate more data while preserving the causality relevant information in the original dataset.

1Cademy (Nik et al., 2022) experimented with self-training to generate more sequence classification examples from unlabeled Wikipedia sentences. They experimented with three pretrained models (BERT, RoBERTa and ELECTRA), and also experimented with three ratios of positive to negative self-labeled examples (1:3, 1:1, 3:1). Their experiments showed that including self-labeled data during training always returns higher F1 scores. Their best model during test time was the RoBERTa-based model with 1:1 self-training ratio, which surpassed the competition baseline scores.

GGNN (Trust et al., 2022) injected word embeddings into a Gated Graph Neural Network (GGNN), which were attached to a RNN decoder to predict the sequence label. Two word embeddings were explored: Word2Vec and BERT. Their BERT+GGNN combination outperforms the BERT baseline provided during the competition for both the development and test sets for P, F1 and Acc.

6.2.2 Subtask 2

1Cademy (Chen et al., 2022) approached this task in a reading comprehension manner, and created a baseline BERT-based neural network that predicted the start and end positions of each Cause, Effect, and Signal span. They introduced beam-search methods (BSS) as post-processing constraints suited to the task. They also introduced a signal classifier that detects if a Signal exists in the sequence or not via a joint model (JS) or a separate model (ES). Additionally, BART was fine-tuned for paraphrasing to re-write Cause and Effect phrases within each sentence for data augmentation (DA). In the end, their best model is a combination of Baseline+BSS+ES+DA method, where the DA generated 3 new phrases per span. This model achieved F1 score of 54.15% on the test set, clinching the top place during the competition.

IDIAPers (Fajcik et al., 2022) approached the task in an encoder-decoder framework. They conditioned the T5 language model three times per example to generate up to four causal relations per example. In each round, given the history of a sentence, the model generates Cause, followed by Effect, and then Signal. This model is their vanilla model known as T5-CES. History refers to the input sequence with any annotated spans from the previous round, if applicable. In experiments, they also explored (1) variants involving a version without historical annotations, (2) T5-large pre-trained

model, and (3) changing the order of generation to be Effect, Cause then Signal. Their best model on the test set (T5-CES) achieved 48.8 F1 score, coming in second in the competition.

SPOCK (Saha et al., 2022) designed two separate frameworks for the span detection task, span-based modelling and token classification. Both approaches far exceed the random baseline provided by the organizers during the competition period. Their span-based modelling approach achieved an F1 score of 47.48%, ranking third in the competition. This model classifies a list of candidate spans to a Cause, Effect, Signal or None label. The candidate spans are generated by considering all possible spans up to a maximum length. The model receives inputs comprising a CLS token embedding, concatenated with a width embedding, plus the span embedding representation itself. To select the final Cause-Effect-Signal span, spans below a certain threshold are removed, and then the span with the highest probability for that label is retained.

LTRC (Adibhatla and Shrivastava, 2022) approached the task as a token classification task, and designed a BERT-based IOB predicting model alongside some heuristics adjusted for the task. Their approach slightly beats the baseline scores on the development set.

7 Analysis & Discussion

7.1 Trends

Consistent with NLP trends, pre-trained language models are popular and employed by all teams and for both subtasks.

For Subtask 1, teams found novel ways to improve from the BERT and LSTM baseline by combining multiple models, adding linguistic features, incorporating additional neural network layers, and working with augmented data.

For Subtask 2, there is a wide variation in framing the task. Teams approached it as a reading comprehension, encoder-decoder, candidate span classification and token classification task. Additionally, there are two constraints that models had to accommodate: (1) The task involves predicting multiple causal relations per input sentence, and (2) Not all causal relations have a signal span. The top three teams carefully adjust their models to work with the two constraints. For (1), IDIAPers predicted different relations using rounds while incorporating the predicted annotations of
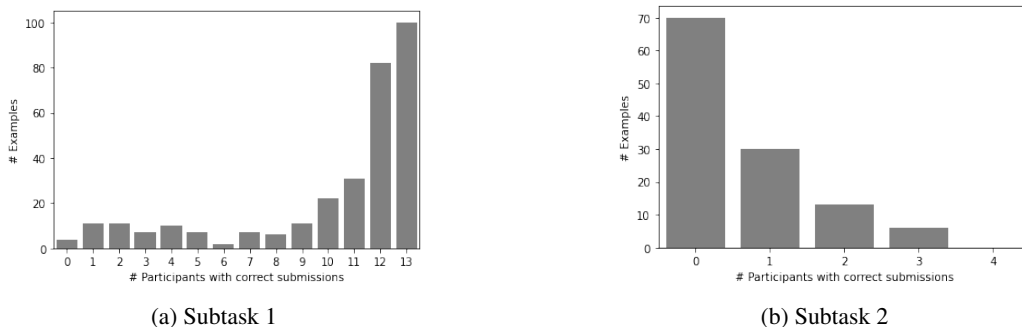
(a) Subtask 1

(b) Subtask 2

Figure 4: Histogram of exact matches.

Subtask	Finished	Failed	Total
Subtask 1	58	8	66
Subtask 2	12	24	36

Table 7: Number of submissions received for test set.

the previous round. For (2), 1Cademy included a separate classification task, while IDIAPers and SPOCK permitted "empty" or "None" span predictions. Interestingly, the F1 score for signals is highest for IDIAPers, suggesting merits to predict signal spans in a manner that includes Cause and Effect predictions as inputs.

7.2 Participation

More submissions were received for Subtask 1 than for Subtask 2, as shown in Table 7. Unsurprisingly, there is a high proportion of failed submissions in Subtask 2. Since Subtask 2 requires specific formatting of argument markings and compiling of multiple predictions into a list, it is easy to face formatting errors. For Subtask 2, although 12 participants did try to submit for the competition, only 4 managed to submit predictions of the right format. A closer look at the submission files suggests that most of the time, these participants intended to upload predictions for Subtask 1. However, because the default Codalab tab falls on Subtask 2, they make submissions to the wrong task. Nevertheless, we are aware of 1 participant who reached out to try and resolve formatting issues and did not manage to resubmit their predictions in the right format in time. This team ran into issues trying to match the spacing of the original input text.

7.3 Error Analysis

For Subtask 1, we had 13 participants while for Subtask 2, we had 4 participants. For Subtask 1, we counted the number of teams that matched the

true labels exactly per example. For Subtask 2, if any predicted span exactly match the true Cause-Effect-Signal span, we considered there to be an accurate count. A histogram per subtask reflecting the accuracy counts are reflected in Figure 4.

For Subtask 1, 100 examples were predicted correctly, while 4 examples were predicted wrongly by all participants. There is a total of 52 examples that are challenging, where less than half of the participants were able to get a correct prediction.

For Subtask 2, no examples were predicted correctly by all participants. This is because LTRC's submission was very close to the Random Baseline and had no exactly correct predictions. 6 causal relations were predicted correctly by the remaining three participants. Nevertheless, most examples were predicted wrongly by all participants (i.e., 70 examples received all wrong predictions). Clearly, Subtask 2 is a challenging task and has a lot of room for growth.

8 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 29 registered participants and 17 active participants who made over 100 submissions on the test set. Based on the 12 description papers received, many novel methods that exceeded our initial baseline were proposed. The best F1 scores achieved for Subtask 1 and 2 were 86.19% and 54.15% respectively.

We intend to re-launch this shared task next year with even more data for Subtask 2. Additionally, we will also investigate the challenging examples in Subtask 1 that are predicted wrongly by many teams.

References

Hiranmai Sri Adibhatla and Manish Shrivastava. 2022. LTRC @ Causal News Corpus 2022: Extracting and identifying causal elements using adapters. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

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

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy. 2022. CSECU-DSG @ Causal News Corpus 2022: Fusion of RoBERTa transformer variants for causal event classification. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

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

Sergio Burdisso, Juan Zuluaga-Gomez, Martin Fajcik, Esaú Villatoro-Tello, Muskaan Singh, Petr Motlicek, and Pavel Smrz. 2022. IDIAPers @ Causal News Corpus 2022: Causal relation identification using a few-shot and prompt-based fine-tuning of language models. In *The 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE @ EMNLP 2022)*, Online. Association for Computational Linguistics.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4862–4872, Online. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. 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 2022)*, Online. 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.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. Association for Computational Linguistics.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan. The COLING 2016 Organizing Committee.

Martin Fajcik, Muskaan Singh, Juan Zuluaga-Gomez, Esaú Villatoro-Tello, Sergio Burdisso, Petr Motlicek, and Pavel Smrz. 2022. IDIAPers @ Causal News Corpus 2022: Extracting cause-effect-signal triplets via pre-trained autoregressive language model. In *The 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE @ EMNLP 2022)*, Online. Association for Computational Linguistics.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. 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 1808–1817, Minneapolis, Minnesota. 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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multiway classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. 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: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Ali Hürriyetoğlu, editor. 2021. *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. Association for Computational Linguistics, Online.

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Farhana Ferdousi Liza, Ritesh Kumar, and Shyam Ratan. 2021a. 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.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, Osman Mutlu, Deniz Yuret, and Aline Villavicencio. 2021b. Challenges and applications of automated extraction of socio-political events from text (CASE 2021): Workshop and shared task report. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*

(CASE 2021), pages 1–9, Online. Association for Computational Linguistics.

Ali Hürriyetoğlu, Erdem Yörük, Vanni Zavarella, and Hristo Tanev, editors. 2020a. *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*. European Language Resources Association (ELRA), Marseille, France.

Ali Hürriyetoğlu, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020b. Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 1–6, Marseille, France. European Language Resources Association (ELRA).

Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Fırat Duruşan, Çağrı Yoltar, Deniz Yüret, and Burak Gürel. 2021. Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction. *Data Intelligence*, 3(2):308–335.

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.

Juhyeon Kim, Yesong Choe, and Sanghack Lee. 2022. SNU-Causality Lab @ Causal News Corpus 2022: Detecting causality by data augmentation via part-of-speech tagging. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

Guneet Kohli, Prabsimran Kaur, and Jatin Bedi. 2022. ARGUABLY @ Causal News Corpus 2022: Contextually augmented language models for event causality identification. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

Theresa Krumbiegel and Sophie Decher. 2022. NLP4ITF @ Causal News Corpus 2022: Leveraging linguistic information for event causality classification. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

Grace E. Lee and Aixin Sun. 2019. A study on agreement in PICO span annotations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1149–1152. ACM.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (FinCausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32, Barcelona, Spain (Online). COLING.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2021. The financial document causality detection shared task (FinCausal 2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 58–60, Lancaster, United Kingdom. Association for Computational Linguistics.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

Quynh Anh Nguyen and Arka Mitra. 2022. NoisyAnnot @ Causal News Corpus 2022: Causality detection using multiple annotation decision. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

Adam Nik, Ge Zhang, Xingran Chen, Mingyu Li, and Jie Fu. 2022. 1Cademy @ Causal News Corpus 2022: Leveraging self-training in causality classification of socio-political event data. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering, Papers from 2003 AAAI Spring Symposium, Stanford University, Stanford, CA, USA*, pages 28–34. AAAI Press.

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*.

Anik Saha, Alex Gittens, Jian Ni, Oktie Hassanzadeh, Bulent Yener, and Kavitha Srinivas. 2022. SPOCK @ Causal News Corpus 2022: Cause-effect-signal span detection using span-based and sequence tagging models. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Roser Saurı, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. Timeml annotation guidelines version 1.2.1.

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.

Fiona Anting Tan, Devamanyu Hazarika, See-Kiong Ng, Soujanya Poria, and Roger Zimmermann. 2021. Causal augmentation for causal sentence classification. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 1–20, Punta Cana, Dominican Republic. 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. 2022a. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng. 2022b. Unicausal: Unified benchmark and model for causal text mining.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Paul Trust, Provia Kadusabe, Rosane Minghim, Ahmed Zahran, Evangelos Milos, Kizito Omala, and Haseeb Yonais. 2022. GGNN @ Causal News Corpus 2022: Gated graph neural networks for causal event classification from social-political news articles. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Erdem Yörük, Ali Hürriyetoğlu, Fırat Duruşan, and Çağrı Yoltar. 2021. Random sampling in corpus design: Cross-context generalizability in automated multicountry protest event collection. *American Behavioral Scientist*, 0(0):00027642211021630.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021a. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172, Online. Association for Computational Linguistics.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021b. LearnDA: Learnable knowledge-guided data augmentation for event causality identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3558–3571, Online. Association for Computational Linguistics.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1544–1550, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Appendix

A.1 Subtask 2 Agreement Score Calculations

For span annotations, the agreement scores were calculated by taking a weighted average of the subset level agreement scores that takes into account the example counts per subset.

We split the training plus development set into 8 subsets and the test set into 2 subsets. While conducting the annotations, the agreement scores were evaluated at a subset level so that we can consistently assess the annotators' performance. The subset level scores takes the average scores between each pair of annotators. For example, if there were three annotators (Annotator A, B, and C) for the subset, then we took the average agreement score when comparing (A,B), (B,C) and (A,C) annotator pairs. Each pair was weighted equally.

The annotator pair level scores were computed by taking the average scores across the sentences. Each sentence was weighted equally.

At the sentence level, agreement scores were obtained by taking the average scores of each causal relation pair. Each causal relation pair was weighted equally.

Since annotators might annotate multiple spans per example, there are many ways to match the annotated relations between two annotators. We approached this conflict by considering every possible combination pair, after which, we retained

the match that returned the highest possible sum of EM, OSB and TO scores. If one annotator identified more causal relations than the other, then EM, OSB and TO scores for that relation is automatically zero.

The KAlpha script was an open-source code[9]. The other three metrics were coded based on previous work (Lee and Sun, 2019).

---

[9]https://github.com/emerging-welfare/kAlpha