RECESS: Resource for Extracting Cause, Effect, and Signal Spans

[1]Fiona Anting Tan, [2]Hansi Hettiarachchi, [3]Ali Hürriyetoğlu, [4]Nelleke Oostdjik, [5]Tommaso Caselli, [6]Tadashi Nomoto, [7]Onur Uca, [8]Farhana Ferdousi Liza, [1]See-Kiong Ng
[1]Institute of Data Science, National University of Singapore, Singapore
[2]School of Computing and Digital Technology, BCU, United Kingdom
[3]KNAW Humanities, Cluster DHLab, The Netherlands
[4]Centre for Language Studies, Radboud University, The Netherlands
[5]CLCG, University of Groningen, The Netherlands
[6]National Institute of Japanese Literature, Japan
[7]Department of Sociology, Mersin University, Turkey
[8]School of Computing Sciences, University of East Anglia, United Kingdom
tan.f@u.nus.edu

Abstract

Causality expresses the relation between two arguments, one of which represents the cause and the other the effect (or consequence). Causal relations are fundamental to human decision making and reasoning, and extracting them from natural language texts is crucial for building effective natural language understanding models. However, the scarcity of annotated corpora for causal relations poses a challenge in the development of such tools. Thus, we created Resource for Extracting Cause, Effect, and Signal Spans (RECESS), a comprehensive corpus annotated for causality at different levels, including Cause, Effect, and Signal spans. The corpus contains 3,767 sentences, of which, 1,982 are causal sentences that contain a total of 2,754 causal relations. We report baseline experiments on two natural language tasks (Causal Sentence Classification, and Cause-Effect-Signal Span Detection), and establish initial benchmarks for future work. We conduct an in-depth analysis of the corpus and the properties of causal relations in text. RECESS is a valuable resource for developing and evaluating causal relation extraction models, benefiting researchers working on topics from information retrieval to natural language understanding and inference.

1 Introduction

A causal relation encodes a semantic relationship between two arguments, where one is the Cause argument, and the other is the Effect argument, in which the occurrence of the Cause leads to the occurrence of the Effect (Barik et al., 2016). A Cause can be a reason, explanation or justification that leads to an Effect (Webber et al., 2019). Causal relation extraction is an important information retrieval (IR) and natural language processing (NLP) task. Research has shown the usefulness of extracting causal relations in text for applications like summarization and next event 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). For example, Izumi et al. (2021) built a prototype using a database of extracted causal relations from financial documents such that users can search for historical causal chains to anticipate upcoming events. Despite the benefits of causal relation extraction, answering *why* something has happened is not a trivial task for multiple reasons, ranging from the fact that a clear definition is needed of what the optimal answer should contain (Dunietz et al., 2020), to the lack of explicit evidence in a text due to a reference to commonsense knowledge. Another difficulty arises because causal relationships may be general or specific (Mackie, 1980; Hitchcock, 1995), and so it is crucial to identify in what context the causal relation occurs. The importance of causal relations to humans, and the difficulty in understanding them, is the main drive of our work to annotate a dedicated dataset for causal text mining. By providing a comprehensive annotated corpus of causal relations, we aim to facilitate the development of more effective causal relation extraction models.

In this paper, we present a Resource for Extract-

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 66–82
November 1–4, 2023.

Sentence	Label	Span Annotations
The bombing created panic among villagers .	*Causal*	<cause>The bombing</cause> <effect><signal>created</signal> panic among villagers</effect> .
Lack of medical services because of the strike left several patients in agony .	*Causal*	<effect>Lack of medical services</effect> <signal>because of</signal> <cause>the strike</cause> left several patients in agony . <cause>Lack of medical services</cause> because of the strike <effect><signal>left</signal> several patients in agony</effect> .
KSRTC buses were attacked at ten places .	*Non-causal*	-

Table 1: Annotating sentences with binary labels, *Causal* or *Non-causal*, and annotating *Causal* sentences with *Cause*, *Effect* and *Signal* spans.

ing Cause, Effect, and Signal Spans (RECESS).[1] The most relevant contribution is the fine-grained annotation of Cause, Effect and Signal spans in causal sentences. Some examples are shown in Table 1. These richer annotations allow us to perform investigations into properties of causal relations in text (See Section 6 for detais). Additionally, we create annotation guidelines and evaluation rules that can accommodate multiple causal relations in one sequence, so as to allow the study of causal chains.

We provide competitive baseline scores based on state-of-the-art models for two NLP tasks: (1) Causal Sentence Classification and (2) Cause-Effect-Signal Span Detection. With a total of 2,574 causal relations, RECESS is larger than other causal text mining benchmarks: CausalTimeBank (CTB) (Mirza et al., 2014) contains 318 causal pairs and EventStoryLine (ESL) (Caselli and Vossen, 2017) contains 1,770 causal pairs. In our experiments, we demonstrate that RECESS' data size is sufficient to train models that return reasonable F1 scores for both tasks. We also propose a rule-based scheme to convert RECESS into SQuAD format, and demonstrate its relevance to Why-Questions. To promote research on causal text mining, we hosted a shared task using RECESS.[2]

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 outlines the creation process of RECESS while Section 4 analyzes the properties of the final dataset. Section 5 presents our experiments, models, and their evaluation against RECESS. Section 6 investigates some research questions about causal relations in text. Finally, Section 7 concludes this paper.

[1]Our repository is available at: https://github.com/tanfiona/CausalNewsCorpus. RECESS is equivalent to CNC-V2.

[2]At the point of this paper's submission, the Event Causality Identification Shared Task (Tan et al., 2023) was in progress.

2 Related Work

Extraction of causality from text is a non-trivial task since the semantic understanding of the context and world knowledge is often necessary. Automatic extraction of causal knowledge from text has been of interest to many computational linguistic researchers (Blanco et al., 2008; Do et al., 2011; Kontos and Sidiropoulou, 1991; Riaz and Girju, 2013). There have been corpora that have been annotated for causal events, like the CTB (Mirza et al., 2014), CaTeRS (Mostafazadeh et al., 2016) and ESL (Caselli and Vossen, 2017). However, they only include annotations of event root words and, thus, do not take into account contexts relevant to the Cause and Effect events. Furthermore, CTB only annotates explicit causal relations. On the other hand, the Penn Discourse Treebank (PDTB) (Prasad et al., 2008; Webber et al., 2019; Prasad et al., 2006) includes many causal relations but only between clauses. Hence, PDTB does not capture relationships of more fine-grained arguments within clauses. Our approach is mostly inspired by the BECauSE 2.0 corpus (Dunietz et al., 2017) to include more varied constructions of causality. However, BECauSE 2.0 only contains 1,803 causal relations, of which, more than half rely on resources that require paid access.

To address the lack of publicly available corpora suitable for causal text mining, the Causal News Corpus (CNC) (Tan et al., 2022b) was created. Subsequently, the authors also annotated a small subset of their data with Cause-Effect-Signal spans for the Event Causality Identification shared task (Tan et al., 2022a). However, their work was incomplete: only 264 sentences were annotated from CNC, and the guidelines, descriptions and analyses were brief. In RECESS, we annotated all of CNC's *Causal* sentences. Having a completed dataset allows us to provide deeper analyses on the CNC

data which previously was not possible, especially about the properties of causal relationships in text, as discussed in Section 6. Furthermore, we also revised the binary labels of sentences for which we did not find causal relations during the span annotation phase, and performed sentence splitting for examples found to contain multiple sentences. Therefore, on top of having more fine-grained span annotations, RECESS is also a larger and cleaner corpus than CNC for binary causal relation classification.

3 The RECESS Corpus

In RECESS, we annotated *Causal* sentences with Cause, Effect and Signal spans, where available.

3.1 Defining Causal Relations

A sentence that contains at least one Cause and Effect pair is said to be *Causal*. The causal relation must be expressed in the target sentence, regardless of its truthfulness in the world.

To define causality more concretely, we used the five tests for causality from Grivaz (2010) to logically verify the Cause and Effect spans, in a similar way to previous works (Dunietz et al., 2017; Tan et al., 2022b,a). The tests are described as follows:

1. **Why:** The event pair is not causal if the reader cannot construct a "Why" question based on the Effect.

2. **Temporal order:** The event pair is not causal if the Cause does not precede the Effect in terms of time.

3. **Counterfactual:** The event pair is not causal if the Effect is equally likely to occur or not occur without the Cause.

4. **Ontological asymmetry:** The event pair is not causal if the reader can swap the Cause and Effect spans.

5. **Linguistic:** The event pair is likely to be causal if it can be rephrased into "*X causes Y*" or "*Because of X, Y.*"

All five checks must be met in order for a pair of events to be considered *Causal*. Annotators used this framework to propose annotations, and Table 2 demonstrates this in application.

3.2 Data Source

RECESS builds on the CNC (Tan et al., 2022b), which is based on the GLOCON dataset of an Event Extraction Shared Task at CASE2021 (Hürriyetoğlu et al., 2021). CNC contains 869 news documents and 3,559 English sentences. The news were reported from the year 2000 to the beginning of 2018 (Hürriyetoglu et al., 2021). After post-curation and cleaning (discussed later), the final dataset comprises of 3,767 sentences. We used the same train-validation-test splits as CNC.

3.3 Annotation Guidelines

We define a Cause or Effect span as a continuous set of words sufficient for interpreting the causal relation. This means that any context modifying or describing the argument relevant to the causal relation is included. Each Cause or Effect span must contain an event, where an event is defined by a situation that happens or occurs, or a predicate that describes a state or circumstance in which something obtains or holds true (Pustejovsky et al., 2003).

Signal spans refer to words that explicitly relate the Cause and the Effect argument. Signals can occur in any position in a sentence relative to the Cause and Effect arguments. Signals can be comprised of multiple words (E.g. *"so intense that"*), and can also be discontinuous (E.g. *"if...then"*), again, consistent with the treatment of connectives by BECAUSE (Dunietz et al., 2017) and PDTB (Prasad et al., 2008; Webber et al., 2019; Prasad et al., 2006). Signals are usually classified into three types: explicit (relations are signaled by discourse connectives), alternative lexicalizations (AltLex) (relations are signaled by a different lexical form), and implicit (no connectives) (Webber et al., 2019; Knaebel and Stede, 2022; Hidey and McKeown, 2016).

In the Appendix A.1, we highlight the exclusion rules, and on how we annotated multiple causal relations. For detailed explanation and examples of the annotation guidelines, please refer to the Annotation Manual available in our repository.

3.4 Annotation and Curation Process

Annotation and curation were conducted using WebAnno (Eckart de Castilho et al., 2016), and post-processed with Python. Appendix A.2 provides a description and screenshots of this tool.

Five annotators, from different academic back-

Sentence	Causality Tests					Label
	Why?	Temporal Order	Counter-fact.	Onto. Asym.	Linguis-tic	
<cause>This strike</cause> <signal>is causing</signal> <effect>huge disruptions</effect>...	✓	✓	✓	✓	✓	*Causal*
<potential-effect>Some protesters attacked me</potential-effect> when <potential-cause>I was clicking pictures<potential-cause>...	✗	✓	✗	✓	✓	*Non-causal*

Table 2: Examples illustrating how to use the Five Tests for Causality to check span annotations.

grounds, were involved in the span annotations. Four curators, comprising of a linguistics undergraduate, linguistics graduate, NLP PhD student, and linguistics professor, were also involved. After each annotation round, a curator considered the spans proposed by all annotators and decided on the final annotations. For some subsets, a second curator looked through these final annotations and discussed disagreements. At the end of every annotation round, the final span annotations were made available to all annotators and curators to review and dispute.

Curators also helped to revise annotations from earlier rounds to adhere to the latest guidelines. Some examples wrongly contained multiple sentences due to the reliance on automatic splitting of sentences from the GLOCON dataset. Thus, curators also helped to mark locations to split up such instances into single sentences. For sentences where no causal relations were found, we revised the sequence classification label to *Non-causal* instead.

3.5 Inter-annotator Agreement

Given two proposed annotations for a causal relation, we first computed the Krippendorf's Alpha (K-Alpha) (Krippendorff, 2011)[3] scores for the Cause, Effect and Signal spans independently. Given a sentence of $n/2$ words, we revised the span annotations into a list of 0s and 1s, where 1 represents that the word is part of the identified span, 0 otherwise. n_0 is the number of words that are not tagged as part of the span in the both annotations, while n_1 is the opposite. This means that $n = n_0 + n_1$ is the total number of words from both annotations. o_{01} is the number of words where the two annotators disagree. K-Alpha is then represented by:

$$\alpha_{binary} = 1 - (n-1) \cdot \frac{o_{01}}{n_0 \cdot n_1} \quad (1)$$

[3]Code to calculate K-Alpha across multiple annotators: https://github.com/emerging-welfare/kAlpha.

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 3: Sequence Labels for Event Sentences 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 4: Span Annotations for Causal Sentences Summary Statistics.

Since annotators can propose multiple causal relations per example, we considered all possible combination to match the proposed relations, and kept the match that returned the highest possible sum of Exact Match (EM), One-Side Bound (OSB) and Token Overlap (TO) scores. The calculations for EM, OSB and TO are detailed in Appendix Section A.3.

To obtain the K-Alpha score per example, we weighted the score of each span by its true length (i.e. number of words). Essentially, the K-Alpha score weights each true word equally.

Finally, we aggregated the scores across examples to obtain the dataset level scores. Again, each example's score was weighted by its true span lengths. Overall, the inter-annotator agreement score (based on K-Alpha (Krippendorff, 2011)) is 42.66%, 44.36% and 28.45% for Cause, Effect and Signal spans respectively, and 42.96% overall.

Topic	Type	Example Annotated Sentence
Arg. Order in Text	*Cause* before *Effect*	<signal>Since</signal> <cause>the work on the bridge was progressing at a snail ' s pace</cause> , <effect>the locals had begun an agitation since June 16</effect> .
	Effect before *Cause*	The R82 , <effect>an arterial road linking Vereeniging to Johannesburg remains closed</effect> <signal>due to</signal> <cause>the public unrest</cause> .
No. and Flow of Events	Single Event	<signal>Since</signal> <cause>the work on the bridge was progressing at a snail ' s pace</cause> , <effect>the locals had begun an agitation since June 16</effect> .
	Concurrent Events	<cause>Police opened fire</cause> , <effect><signal>killing</signal> 34 striking workers and <signal>wounding</signal> 78</effect> ...
	Consecutive Events	<effect>Lack of medical services</effect> <signal>because of</signal> <cause>the strike</cause> left several patients in agony . <cause>Lack of medical services</cause> because of the strike <effect><signal>left</signal> several patients in agony</effect> .
Signal	Explicit	The R82 , <effect>an arterial road linking Vereeniging to Johannesburg remains closed</effect> <signal>due to</signal> <cause>the public unrest</cause> .
	AltLex	<cause><signal>Irked over</signal> the arrests</cause> , <effect>the protestors staged dharna</effect> .
	Implicit	However, <cause>trade unions refused to accept it</cause> and <signal>*(Implicit=thus)*</signal> <effect>continued their strike</effect>.
Sense	Cause	<effect>Spokesman Keith Khoza said they had decided to March to Prime Media</effect> <signal>because</signal> <cause>the cartoon had raised various concerns</cause>.
	Purpose	<effect>The protesters planted the saplings in potholes</effect> <cause><signal>to</signal> draw the attention of the officials to the poor condition of the road</cause> .
	Condition	<signal>If</signal> <cause>it does not act</cause>, <effect>the protests will continue</effect> .
	NegCondition or NegResult	<effect>We will continue our strike</effect> <signal>till</signal> <cause>we get an assurance from the Government</cause> .

Table 5: Examples from RECESS.

Rank	Signal	Count	Rank	Signal	Count
1	to	364	11	because	20
2	for	170	12	in protest	19
3	as	141	13	in connection with	17
4	demanding	111	14	in support of	14
5	following	78	15	seeking	12
6	by	64	16	even as	12
7	over	64	17	protesting	10
8	if	46	18	despite	10
9	with	44	19	in the wake of	10
10	due to	25	20	led to	10

Table 6: Top 20 Signals in RECESS.

4 Dataset Analysis

Table 3 and 4 present summary statistics of RECESS. RECESS contains 1,982 *Causal* and 1,785 *Non-causal* sentences. In total, the *Causal* sentences correspond to 2,754 causal relations. This means that each sentence has an average number of 1.39 causal relations. The distribution of the number of causal relations in the corpus is as follows: 1,354 sentences have one causal relation, while 498, 118, 10, and 2 sentences have 2, 3, 4 and 5 causal relations, respectively.

In RECESS, *Causal* sentences are longer than *Non-causal* sentences in terms of word counts. More specifically, *Causal* sentences have an average length of 33.75 words while *Non-causal* sentences have an average length of 26.90 words. Finally, since Signals can be implicit, the average number of relations that have Signals marked is less than one, at 0.69.

Table 5 presents some annotated sentences while Table 6 highlights the 20 most common Signals from RECESS. The annotation guidelines support a wide array of linguistic, syntactic and semantic structures in terms of (1) argument order in text, (2) number and flow of events, (3) signal type, and (4) sense type. Therefore, RECESS is a useful, diverse corpus for various types of linguistic and NLP research.

5 Experiments

In this Section, we demonstrate use-cases of RECESS for NLP research.

5.1 Tasks

RECESS is suitable for Event Causality Identification tasks, which aim to design models that tackle problems such as:

1. Causal Sentence Classification (CSC): Does an event sentence contain any cause-effect meaning?

2. Cause-Effect-Signal Span Detection (CES-SD): Which spans correspond to cause, effect or signal per causal sentence?

5.2 Models

For the CSC task, we replicate CNC's BERT benchmark (Tan et al., 2022b). The model fine-tunes the pre-trained (PTM) BERT model (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 experimented with both `bert-base-cased` and `bert-large-cased`.

For the CES-SD task, we replicate the winning submission (Chen et al., 2022)[4] for the CES-SD Event Causality Identification Shared Task (Tan et al., 2022a). Chen et al. 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. On top of this baseline, they developed three components to further improve model performance: Beam-search span selector (BSS), signal classifier (SC), and data augmentation (DA). We incrementally incorporated these components on top of **Baseline**, resulting in the three additional models investigated: **Baseline+BSS**, **Baseline+BSS+SC** and **Baseline+BSS+SC+DA**. All models fine-tune `albert-xxlarge-v2` (Lan et al., 2019). We describe the model and components in detail in the Appendix B. Model hyperparameters are available in the Appendix C.

5.3 Experimental Setup

We trained each model on the training set and used the development set to select the best model with the highest F1 score at the end of each epoch. Subsequently, this best model was used to predict the test set.

5.4 Evaluation Metrics

For CSC, we evaluate using Accuracy (Acc), standard Precision (P), Recall (R) and F1 per class, and Matthews Correlation Coefficient (MCC).

For CES-SD, we evaluate using token-wise Macro Precision (P), Recall (R) and F1 metrics. We used the FairEval implementation[5] of `seqeval` (Nakayama, 2018; Ramshaw and Marcus, 1995) that prevents double penalties of close-to-correct predictions (Ortmann, 2022). We evaluate at the relation level: each relation contributed equally to the final score. For sentences with multiple causal relations, as in Tan et al. (2022a), we returned the highest F1 score out of every possible way to match the predicted and true causal relations to each other. We only compare predictions with the number of true causal relations available. Conversely, any missing predictions were interpreted to predict the Other (`O`) label for all tokens.

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

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

Eval	PTM	R	P	F1	Acc	MCC
Dev	base	**88.65**	84.10	**86.32**	**84.71**	**69.13**
	large	84.86	**85.79**	85.33	84.12	68.02
Test	base	**89.02**	75.86	81.91	80.68	62.37
	large	88.44	**78.46**	**83.15**	**82.39**	**65.35**

Table 7: Performance Metrics for Causal Sentence Classification. All scores are reported in percentages (%). Highest score per dataset and metric is in bold.

5.5 Baseline Scores

From Table 7, for the CSC task, the `base` BERT variant achieved an F1 score of 86.32% for the development set. Although the `large` BERT variant performed worse than `base` for the development set, it performed much better than `base` for the test set, achieving an F1 score of 83.15%.

From Table 8, for the CES-SD task, the best model is **Baseline+BSS+SC**, with a score of 70.51% for the development set and 67.69% for the test set. For Chen et al. (2022), **Baseline+BSS+SC+DA** was the best performing model. DA might have been useful for them because their training data size was small, with only 264 sentences. For us, incorporating artificially augmented data did not improve performance. This might be because there is limited linguistic diversity in the augmented data, since the DA approach fixed the signals and only paraphrased the Cause and Effect arguments. Furthermore, upon investigation, many of the sentences are grammatically unsound. For more analyses about the Subtask 2 model variants, please refer to Appendix B.

All in all, we report scores that will serve as initial baselines for future researchers to beat. Coupled with strong CSC and CES-SD models, researchers can perform end-to-end extraction of causal relations from text.

Eval	Model	Overall		
		R	P	F1
Dev	**Baseline**	66.32	59.48	62.71
	+BSS	**71.39**	64.43	67.73
	+BSS+SC	71.22	**69.81**	**70.51**
	+BSS+SC+DA	70.89	69.25	70.06
Test	**Baseline**	61.49	61.89	61.69
	+BSS	**67.30**	66.98	67.14
	+BSS+SC	66.56	**68.86**	**67.69**
	+BSS+SC+DA	64.43	67.56	65.96

Table 8: Performance Metrics for Cause-Effect-Signal Span Detection. All scores are reported in percentages (%). Highest score per dataset and metric is in bold. Detailed results available at Appendix Table 12.

6 Investigations

In this section, we explore research questions about causal relations in text using RECESS' training and development set. We intend to keep the test set as an unseen, hold-out set to be used for future shared tasks, and thus, do not perform analyses on it.

6.1 When is causality easy/hard to detect?

First, we investigate the identification of causality in text by humans. We categorize sentence classification instances into two types: (1) Annotators all agree with one another, and (2) At least one annotator disagrees with another. From Figure 1a, it is more likely for annotators to all agree that a sentence is *Causal* compared to when it is *Non-causal*. The average observed agreement score of *Causal* sentences is 87.10%, but only 78.23% for *Non-causal* sentences. From Figure 1b, causal relations are easier for humans to detect if there are causal markers present in the text. For *Causal* examples with explicit or AltLex signals, annotators fully agree with each other around 69% of the time. For implicit *Causal* examples, annotators fully agree with each other less frequently – around 55% of the time.

Next, we perform error analysis on the `base` CSC model. For the *Causal* examples of the development set, 122 had explicit/AltLex markers while 63 were implicitly expressed. The model failed to identify $10/122 \approx 8\%$ explicit/AltLex and $11/63 \approx 17\%$ implicit examples. Similar to humans, the model finds it harder to identify causality if relations are expressed implicitly.

We analyse CES-SD using the best model, **Baseline+BSS+SC**. While it managed to score 70.51% for Overall F1, when we focus only on examples with multiple causal relations, the F1 falls to 53.97%. This is because the current model is not properly designed to detect multiple causal relations in a sentence. The models by (Chen et al., 2022) can only predict one Signal span per causal relation, and the multiple Cause and Effect spans obtained from BSS are still with reference to the same Signal. Therefore, the variation in Cause and Effect arises from having only slightly different word boundaries, instead of having different semantic arguments. We hope future researchers will improve on this aspect.

Out of the 249 causal relations in the development set, the model managed to predict 98 sequences exactly as the gold label. Of these, 62 contain explicit/AltLex markers, while the remaining 36 were implicitly expressed. Consistent with earlier findings, it is easier for models to detect causality if explicit/AltLex markers are present.

Models perform well for cases where the Cause and Effect spans are located near to each other. Out of the 62 explicit/AltLex marked examples, 35 had their causal signals located between the Cause and Effect arguments. The remaining examples either had the arguments following one another, if not, separated only by a comma. Likewise, for the 36 implicitly expressed examples, 26 had one-worded discourse markers lying between the Cause and Effect spans. The markers are non-causal ones, like *"after"*, *"when"* and *"in"*. Again, the remaining 10 had sequential arguments, or at most, had arguments separated by a comma.

6.2 Do signals matter?

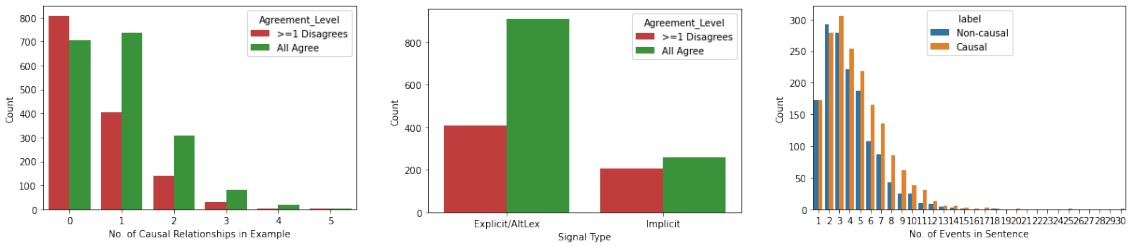
In the investigations above, the presence of causal signals helps indicate the presence of causality. In this subsection, we perform a qualitative study by using the CSC and CES-SD models to predict on six examples in Table 9. We notice that the models are very sensitive to explicit causal markers (E.g. *"because"*) and non-causal markers (E.g. *"but"*) to the extent where the content of the sentence no longer matters. For the examples without linguistic markers (E.g. S/N 2 and 3), the content of the sentence matters.

6.3 How does the presence of causality correlate with the number of events?

The Event Extraction Shared Task (Hürriyetoğlu et al., 2021) provides annotations of event spans with seven labels: `<target>`, `<place>`, `<etime>`, `<organizer>`, `<participant>`, `<trigger>` and

S/N	Text	Predictions		Remarks
		Label	Span	
1	The protest was becoming overheated, thus, the police rushed down onsite.	*Causal*	<cause>The protest was becoming overheated,</cause><signal>thus,</signal><effect>the police rushed down onsite.</effect>	Explicit causal
2	The protest was becoming overheated, the police rushed down onsite.	*Causal*	<cause>The protest was becoming overheated,</cause><effect>the police rushed down onsite.</effect>	Implicit causal
3	The protest was becoming overheated, the police said they were aware.	*Non-causal*	-	Non-causal
4	The protest was becoming overheated, but the police rushed down onsite.	*Non-causal*	-	Illogical - With explicit non-causal marker *"but"*
5	The protest was becoming overheated, thus, the protestors were calm.	*Causal*	<cause>The protest was becoming overheated,</cause><signal>thus,</signal><effect>the protestors were calm.</effect>	Illogical - With explicit causal marker *"thus"*
6	Because fire extinguishes water, pigs can fly.	*Causal*	<signal>Because</signal> <cause>fire extinguishes water,</cause><effect>pigs can fly.</effect>	Illogical - With explicit causal marker *"because"*

Table 9: Ablation study: End-to-end predictions on example sentences. "Illogical" reflects Cause and Effect pairs that are not realistic according to world knowledge and commonsense.



(a) No. of causal relations. (b) Signal type. (c) No. of events.

Figure 1: Count plots of train + development set.

Model	SQuAD Dev			
	All (n=10,655)		Why (n=335)	
	EM	F1	EM	F1
No Pre-training	66.11	72.02	53.43	63.21
Pre-training w/ RECESS	66.59	72.51	55.52	65.16

Table 10: QA Performance.

<fname>. An example event sentence annotation is: "<target>*KSRTC*</target> *buses were* <trigger>*attacked*</trigger> *at ten places* ." Only events under the scope of contentious politics were annotated. In other words, they do not follow the definition of events that we use in Section 3.3 based on Pustejovsky et al.. The scripts to add the event annotations to RECESS' examples are provided in our repository.

Based on Figure 1c, sentences are more likely to be *Causal* if they contain more contentious events. On average, *Causal* sentences contain 4.58 contentious events while *Non-causal* sentences contain 3.96 contentious events. Since *Causal* sentences have to contain at least a pair of Cause and Effect event, while *Non-causal* sentences have no such restriction, it makes sense that *Causal* sentences, on average, contain more events. This could also explain why *Causal* sentences tend to be longer in Table 3, since *Causal* sentences need to describe at least two events. However, better event annotations are needed for better analyses of causality correlating with events.

6.4 How are causal relations related to causal question answering (QA)?

In extractive QA, given a question and the context document, the aim is to find the words that form the answer. SQuAD (Rajpurkar et al., 2016, 2018) is a popular benchmark for QA. A natural way to convert causal relations to suit QA is to form a question using the Cause or Effect, then create the answer using the corresponding argument. This can be done by using templates, such as asking *"Why did* <effect>*?"* and using the Cause span as the answer. Each causal relation in the train set returned two QA examples, created by randomly selecting two out of six templates. Therefore, we obtained 4,514 Why-Questions. The templates and conversion scripts are available in our repository.

Our baseline model is the t5-small (Raffel et al., 2020) model trained on the train set and evaluated on the development set of SQuAD. The improved version first pre-trains on our Why-questions before following the same training sched-

ule. In evaluation, we identify 335 questions that contain the words *'reason'*, *'why'*, *'cause'*, or *'resulted'* and refer to them as Why-Questions. The results are available in Table 10. While the original model achieved 53.42 EM and 63.21 F1, the model with additional pre-training on RECESS achieved a superior score of 55.52 EM and 65.61 F1 for Why-Questions. This exercise shows that RECESS has potential applications for QA too, especially surrounding Why-Questions.

6.5 How are causal relations related to natural language inference (NLI)?

Benchmarks like SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018) and ANLI (Nie et al., 2020) classify a premise and hypothesis to one of the three class labels: `neutral`, `entailment` or `contradict`. One way of using RECESS for an NLI task is to rephrase either the Cause or Effect span as the premise, while the corresponding span will be the hypothesis. The label will certainly not be `contradict`. Typically, NLI treats causal relations as `neutral` events. For example, in *"The farmworkers' demands were not met. The farmworkers' strike resumed on Tuesday."*, a competitive model by Nie et al.[6] classifies this as `neutral` with 99.6% probability. Many possible outcomes can arise when farmworkers' demands were not met. For example, they could have initiated an online campaign, or stormed the capital. Since `entailment` requires that the hypothesis is " definitely correct about the situation or event" in the premise (Williams et al., 2018), most of the time, causal relations are considered `neutral` in NLI.

7 Conclusion

Our paper introduces RECESS, a corpus consisting of 3,767 sentences, among which 1,982 are causal sentences containing a total of 2,754 causal relations. We detail the guidelines and process we used to annotate the Cause, Effect, and Signal spans, covering a broad range of linguistic, semantic, and syntactic structures. Additionally, we benchmarked our baseline models, which achieved competitive scores, with F1 scores of 83.15% and 67.69% on test sets for the CSC and CES-SD tasks, respectively. Our work also investigated causal relations in text and explored the relevance of causal relations to QA and NLI applications.

RECESS is an large data that offers opportunities to design specialized models that can extract multiple causal relations. To encourage research in this direction, we are organizing a shared task using RECESS.

Finally, RECESS is a valuable resource for studying NLP topics related to storyline and causal event chains. Researchers can use the contentious political event span annotations marked by earlier works (Hürriyetoğlu et al., 2021) to investigate events in the text. With its extensive coverage of causal relations, we also demonstrated that RECESS can be adapted into a causal QA dataset.

8 Acknowledgments

This project is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

9 Limitations & Ethics Statement

The main limitation of RECESS and our models is that we only focus on extracting causal relations from text, without checking if the relations hold in the real-world. For example, in Table 9, the CES-SD model will identify causal relations for illogical statements with explicit causal markers. In application, users must perform fact-checking of the extracted causal relations, and/or only extract causal relations from reliable sources, to avoid drawing the wrong conclusions.

All annotators and curators voluntarily participated in the creation of the corpus and are the authors of this paper.

[6] https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

References

Biswanath Barik, Erwin Marsi, and Pinar Öztürk. 2016. Event Causality Extraction from Natural Science Literature. *Res. Comput. Sci.* 117 (2016), 97–107. http://rcs.cic.ipn.mx/2016_117/Event%20Causality%20Extraction%20from%20Natural%20Science%20Literature.pdf

Eduardo Blanco, Nuria Castell, and Dan Moldovan. 2008. Causal relation extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large anno-

tated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 632–642. https://doi.org/10.18653/v1/D15-1075

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*. Association for Computational Linguistics, Vancouver, Canada, 77–86. https://doi.org/10.18653/v1/W17-2711

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)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 100–105. https://aclanthology.org/2022.case-1.14

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*. Association for Computational Linguistics, Online, 70–80. https://doi.org/10.18653/v1/2021.deelio-1.8

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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 294–303.

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*. Association for Computational Linguistics, Online, 7839–7859. https://doi.org/10.18653/v1/2020.acl-main.701

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*. Association for Computational Linguistics, Valencia, Spain, 95–104. https://doi.org/10.18653/v1/W17-0812

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)*. The COLING 2016 Organizing Committee, Osaka, Japan, 76–84. https://www.aclweb.org/anthology/W16-4011

Cécile Grivaz. 2010. Human Judgements on Causation in French Texts. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/145_Paper.pdf

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)*. Association for Computational Linguistics, Baltimore, Maryland, 987–997. https://doi.org/10.3115/v1/P14-1093

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*. International Joint Conferences on Artificial Intelligence Organization, 5003–5009. https://doi.org/10.24963/ijcai.2019/695

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)*. Association for Computational Linguistics, Berlin, Germany, 1424–1433. https://doi.org/10.18653/v1/P16-1135

Christopher Read Hitchcock. 1995. The mishap at Reichenbach fall: Singular vs. general causation. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 78, 3 (1995), 257–291.

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)*. Association for Computational Linguistics, Online, 79–91. https://doi.org/10.18653/v1/2021.case-1.11

Ali Hürriyetoglu, Erdem Yörük, Osman Mutlu, Firat Durusan, Çagri Yoltar, Deniz Yüret, and Burak Gürel. 2021. Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction. *Data Intell.* 3, 2 (2021), 308–335. https://doi.org/10.1162/dint_a_00092

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*. Association for Computational Linguistics, Lancaster, United Kingdom, 19–25. https://aclanthology.org/2021.fnp-1.3

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 (2021), 721–739. https://transacl.org/ojs/index.php/tacl/article/view/2717

René Knaebel and Manfred Stede. 2022. Towards Identifying Alternative-Lexicalization Signals of Discourse Relations. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 837–850. https://aclanthology.org/2022.coling-1.70

John Kontos and Maria Sidiropoulou. 1991. On the acquisition of causal knowledge from scientific texts with attribute grammars. *International Journal of Applied Expert Systems* 4, 1 (1991), 31–48.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

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 (2019). arXiv:1909.11942 http://arxiv.org/abs/1909.11942

David Lewis. 1974. Causation. *The journal of philosophy* 70, 17 (1974), 556–567.

John Leslie Mackie. 1980. *The cement of the universe: A study of causation*. Clarendon Press.

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)*. Association for Computational Linguistics, Gothenburg, Sweden, 10–19. https://doi.org/10.3115/v1/W14-0702

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures. In *Proceedings of the Fourth Workshop on Events*. Association for Computational Linguistics, San Diego, California, 51–61. https://doi.org/10.18653/v1/W16-1007

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

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4885–4901. https://doi.org/10.18653/v1/2020.acl-main.441

Katrin Ortmann. 2022. Fine-Grained Error Analysis and Fair Evaluation of Labeled Spans. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1400–1407. https://aclanthology.org/2022.lrec-1.150

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2008/summaries/754.html

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie L Webber. 2006. The penn discourse treebank 1.0 annotation manual. *IRCS Technical Reports Series* (2006).

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*, Mark T. Maybury (Ed.). AAAI Press, 28–34.

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*, Alain Mille, Fabien Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab (Eds.). ACM, 909–918. https://doi.org/10.1145/2187836.2187958

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*, Stefano Leonardi, Alessandro Panconesi, Paolo Ferragina, and Aristides Gionis (Eds.). ACM, 255–264. https://doi.org/10.1145/2433396.2433431

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text

Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*. https://aclanthology.org/W95-0107

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*. 21–30.

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*. Association for Computational Linguistics, Online, 158–170. https://aclanthology.org/2021.bea-1.17

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)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 195–208. https://aclanthology.org/2022.case-1.28

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Nelleke Oostdijk, Onur Uca, Surendrabikram Thapa, and Farhana Ferdousi Liza. 2023. Event Causality Identification - Shared Task 3, CASE 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*. Association for Computational Linguistics, Varna, Bulgaria (Hybrid).

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*. European Language Resources Association, Marseille, France, 2298–2310. https://aclanthology.org/2022.lrec-1.246

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 11328–11339. http://proceedings.mlr.press/v119/zhang20ae.html

A Dataset

A.1 Guidelines

The Annotation Manual provides definitions, examples, and exclusions in better detail. This Section merely highlights important elements of the Manual and does not serve to be an exhaustive outline of the Manual.

A.1.1 Exclusions

Sentences described by any of the following conditions were excluded from RECESS:

1. No causal relation: Although all examples faced during annotation were already filtered to contain causal relations, this exclusion condition holds as a general rule.

2. Correlation relation: Some relationship exists, but it is unclear what the direction of the relationship is. (E.g. "*Smoking is linked to cancer.*").

3. The Cause or Effect span does not contain an event. (E.g. "*John caused the fire.*")

4. The Cause or Effect span requires discontinuous spans.

5. The Cause and Effect spans must overlap.

A.1.2 Multi-relation Annotation

In RECESS, each sentence may have multiple causal relations, which will all be annotated. The annotation tool and evaluation scheme were designed to support such scenarios.

A consequence of annotating multiple causal relations per sentence is that we are able to detect consecutive events, if any. In cases where a span could be split into subparts that describe a series of Cause and Effect events, annotators were instructed to annotate them as separate spans. This is because we are interested to recreate a storyline, and such examples are important to test if models can truly understand narratives, and identify a series of events where one event leads to another.

In Example 1., there is a series of causal events, where → represents causation: "*an altercation*" <EventA> → "*a youth . . . was allegedly severely injured in a thrashing ...*" <EventB> → "*the clashes erupted* " <EventC>. Instead of combining <EventA> and <EventB> into a single Cause span for 1.(a), we asked annotators to mark only the most recent event as the Cause span.

1. (a) <effect>The clashes erupted</effect> after <cause>a youth belonging to a minority Muslim sect was allegedly severely injured in a thrashing by a youth from the majority sect</cause> following an altercation.

   (b) The clashes erupted after <effect>a youth belonging to a minority Muslim sect was allegedly severely injured in a thrashing by a youth from the majority sect</effect> <signal>following</signal> <cause>an altercation</cause>.

One interesting observation is that for consecutive events where one causes the other (E.g. <EventA> → <EventB> and <EventB> → <EventC>), it is also possible to claim that the union of two previous events (<EventA>+<EventB>) causes the third event: <EventA>+<EventB> → <EventC>. This is consistent with Lewis (Lewis, 1974)'s claims that causation is a transitive relation. Take for example the second sentence shown in Table 1: If asked "*Why were several patients left in agony* EventA*?*", notice that we can answer this question with either "*Because of the lack of medical services* <EventB>." or "*Because of the lack of medical services* <EventB> *because of the strike* <EventA>.". A similar exercise can be done for Example 1. by asking "*Why did the clashes erupt?*". Annotators used this transitive property to check if spans contained sub-events that can be broken down to form consecutive causal events.

A.2 Annotation Tool

Annotation and curation were conducted on WebAnno (Eckart de Castilho et al., 2016).

Figure 2 shows the annotation interface used by annotators. On WebAnno, annotators first highlighted the words corresponding to a span, then indicated if the span is a Cause, Effect or Signal. Subsequently, to reflect the causal relations, annotators linked the spans together by directing the Cause to its Effect span, and if present, the Signal to the same Effect span. Annotations were then downloaded as JSON files and automatically validated for avoidable human errors, such as invalid links (E.g. An Effect points to another Effect) or missing links (E.g. An Effect has no Cause). If such errors were present, an error report was produced and sent to annotators to consider correcting their annotations. Finally, the curator assesses all

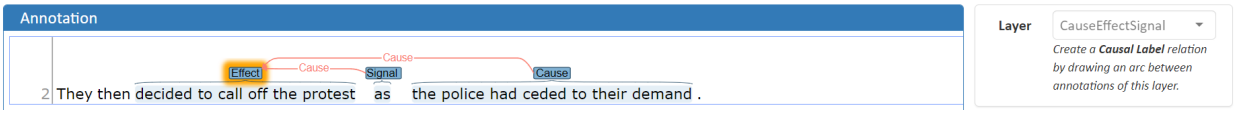


Figure 2: Screenshot of the annotation page used to mark Cause-Effect-Signal spans.

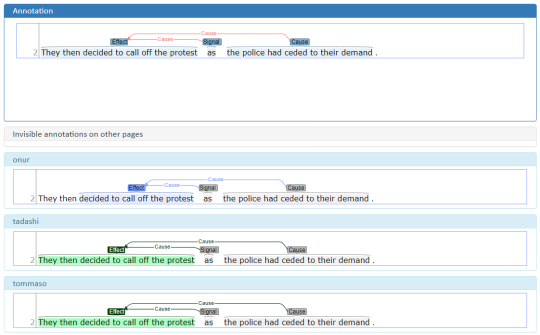


Figure 3: Screenshot of the curation page used to mark Cause-Effect-Signal spans.

the annotations proposed and makes the final selection.

Figure 3 shows the curation interface used by curators. Curators have access to all annotators' annotations, and decide the final annotations.

A.3 Inter-annotator Agreement

Calculating Agreement Scores Given two proposed annotations for a causal relation, we first computed agreement scores for the Cause, Effect and Signal spans individually. On top of the Krippendorf's Alpha (K-Alpha) (Krippendorff, 2011) score described in Section 3.5, we also computed three other agreement metrics as follows:

- Exact Match (EM): 1 if two annotations are exactly the same, 0 otherwise
- One-Side Bound (OSB): 1 if either boundary of the span is the same between two annotations, 0 otherwise
- Token Overlap (TO): 1 if two annotations have at least one overlapped token, 0 otherwise

Since annotators can propose multiple causal relations per example, we considered all possible combination of ways to match the proposed relations. After which, we retained the match that returned the highest possible sum of EM, OSB and TO scores. For example, if one annotator proposed two relations (A,B) while the next annotator proposed three relations (C), then we assessed the scores for (A-C, B-None) and (A-None, B-C) matches, and finally only kept the match that gives the highest sum of scores. If one annotator identified more causal relations than the other, then EM, OSB and TO scores for that relation is automatically zero.

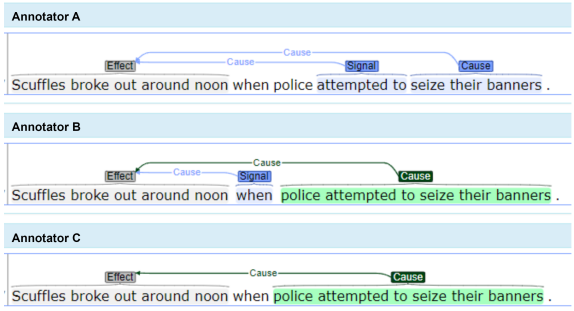
To compute the *Total* score for each relation, EM, OSB, and TO was marked with 1 if and only if the Cause, Effect and Signal all had 1s, otherwise, the example would have an *Total* score of 0.

Finally, we aggregated the scores across examples to obtain the dataset level scores presented in Table 11. For the overall EM, OSB and TO scores, we take the average score across the dataset by weighting each example equally.

In Figure 4, we illustrate one sentence with proposed annotations by three annotators. Since every annotator identified the same Effect span, all four metrics have the score of 1, suggesting full agreement. Meanwhile, since every annotator had a different Signal span, the score for EM, OSB and TO is 0. K-Alpha is negative, suggesting there is less agreement than one would expect by chance. Two annotators had the same proposed Cause span, therefore, EM is equivalent to 1/3, representing 1 out of the 3 possible annotator pairs have exactly matched spans. The Cause span has a score of 1 for OSB and TO, because all three annotators are aligned on the "*banners*" being the last word and part of Cause.

Evaluating Agreement Annotating causal arguments is very challenging. When comparing annotations across two annotators, the statistic is that they will only agree exactly with one another 6.03% of the time. Nevertheless, they will agree with each other 32.51% of the time based on having at least one same token in all three Cause, Effect and Signal arguments. In fact, they do agree with each other's Cause and Effect spans around 57-58% of the time based on having at least one same token. To summarize our annotation experience, it is hard for annotators to agree on both boundaries of a span, perhaps due to the subjectivity of what information is relevant to the event. However, they typically can identify the main event and do include them in

(I) Example Annotations by 3 Annotators



(II) Agreement Metrics for Example

Span Type	Agreement Metric			
	K-Alpha	EM	OSB	TO
Cause	60.00	33.33	100.00	100.00
Effect	100.00	100.00	100.00	100.00
Signal	-26.67	0.00	0.00	0.00
Total	70.44	0.00	0.00	0.00

Figure 4: Example with three annotators, and its corresponding agreement scores reported in percentages (%).

Metric	Span	Train+Dev	Test	Total
K-Alpha	Cause	43.16	38.13	42.66
	Effect	45.01	38.23	44.36
	Signal	29.47	18.85	28.45
	Total	43.55	37.44	42.96
Exact Match	Cause	17.01	13.80	16.69
	Effect	25.78	21.14	25.32
	Signal	34.64	29.03	34.08
	Total	6.03	6.07	6.03
One-Side Bound	Cause	47.68	37.89	46.71
	Effect	50.69	44.17	50.05
	Signal	38.76	34.11	38.30
	Total	26.28	20.95	25.75
Token Overlap	Cause	58.95	49.81	58.04
	Effect	58.00	49.62	57.17
	Signal	39.18	34.16	38.68
	Total	33.07	27.46	32.51

Table 11: Inter-annotator Agreement Scores. Reported in percentages (%).

the annotations. Through our discussions, we also found it challenging to create clear definitions of how to identify a boundary. We made do with the general rule of including words that are: (1) sufficient for interpreting the causal relation, (2) does not change the intended causal relation meaning and (3) up to natural stopping points represented by punctuation marks.

B CES-SD Model

We benchmark four different models based on the winning solution (Chen et al., 2022)[7] of the CASE 2022 CES-SD Shared Task (Tan et al., 2022a) by training and testing them on RECESS. The models frame the task as a reading comprehension task that aims to predict the start and end token positions of each Cause, Effect, and Signal span. All models fine-tune the `albert-xxlarge-v2` (Lan et al., 2019) PLM. More details about the model architectures are described below.

Consider a sentence T with n tokens, such that $T = [t_1, t_2, ..., t_n]$, where t_i represents a token.

Baseline utilizes a BERT-based encoder to convert T into a contextualized representation $R = [r_1, r_2, ..., r_n]$. Each r_i has a depth d representing the hidden size of the PLM, so R has a dimension of $n \times d$. The sequence output R is then fed through a dropout layer followed by a linear layer plus a softmax layer that serves as a classifier. The classifier returns a logit representation, $P = [p_{cs}, p_{ce}, p_{es}, p_{ee}, p_{ss}, p_{se}]$ with a dimension of $n \times 6$. Each p_b vector reflects the probability of each token being the span boundary position. Subscripts cs, ce, es, ee, ss and se refer to Cause-Start, Cause-End, Effect-Start, Effect-End, Signal-Start and Signal-End, respectively. The position with maximum probability was selected as the final prediction using the following formula, where B represents the final position boundary predicted given the logits in p_b:

$$B = \underset{1 \le i \le n}{\operatorname{argmax}} p_b \quad (2)$$

We incrementally incorporated three components on top of **Baseline** as proposed in (Chen et al., 2022), resulting in the three additional models investigated: **Baseline+BSS**, **Baseline+BSS+SC** and **Baseline+BSS+SC+DA**. The components are described below in the next three paragraphs:

BSS Beam-search span selector (BSS) is a post-processing technique used to introduce constraints suited to the task such that (1) the start position is always before an end position, (2) the predicted Cause and Effect spans do not overlap each other, and (3) it is able to return multiple Cause and Effect spans per input sentence. The detailed pseudo-code is presented in the original paper (Chen et al., 2022).

SC Signal classifier (SC) is a separate model that fine-tunes the `bert-base-uncased` (Devlin et al.,

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

2019) PLM on the binary classification task that detects if a Signal exists in a sequence. If the signal classifier model predicts that a sentence contains a Signal, then the Signal predictions from the span detection model will be retained.

DA Data augmentation (DA) was used to generate additional training examples. Cause and Effect spans within each sentence were paraphrased using a PEGASUS model (Zhang et al., 2020). 2257 additional training examples were generated based on the train set. This augmented data is available in our repository.[8]

B.1 Results & Discussion

Due to space limitations in the main paper, we deep dive into the performance for the different CES-SD model variants in this section.

The findings suggest that BSS is a beneficial component to have for CES-SD. Intuitively, BSS constraints the predicted Cause and Effect spans to the task. Therefore, it helps to return a higher score relative to Cause and Effect span predictions. Notice that Signal span predictions, and hence the corresponding performance metrics, are the same compared to the **Baseline**.

For both datasets, performance metrics of **Baseline+BSS** and **Baseline+BSS+SC** have the same scores for Cause and Effect spans. This observation is expected since the SC only affects Signal span predictions by removing them when the classifier identifies the signal to be missing.

Baseline+BSS+SC+DA was the best performing model for (Chen et al., 2022). In their work, the model performance improves when the model is trained on a large set of training data. However, we could not replicate this finding. When we trained the model with augmented data, F1 score fell slightly (from 70.51% to 70.06%) on development set. The test set also observed a drop in performance. When we trained the model on even more augmented data (9 augments per original example), the development set F1 fell even further to 69.28%. Our hypothesis is that the artificially augmented data does not add much linguistic diversity to the training examples. This is because the DA approach fixes the signals and only paraphrases the Cause and Effect arguments. Furthermore, many of the generated sentences are grammatically unsound, for example: "<cause>*There was a lot of violence there*</cause> *has* <effect>*fears of more attacks and heightened tensions*</effect>." In essence, DA corrupts the model by exposing the model with examples that are repetitive and can contain errors. DA might have been useful for (Chen et al., 2022) because the training set back then was extremely small.

[8]We also explored augmenting up to nine additional examples per original example. By obtaining three new phrases per span, we could combine spans such that each causal relation could return nine new examples for training. However, we found poorer results with larger augments in our experiments. The augmented datasets are available in our repository.

C Hyper-parameters

In this Section, we provide the hyper-parameters for each training set up.

For the CSC model, both `bert-base-cased` and `bert-large-cased` variants had the following parameters:

```
per_device_train_batch_size=32
num_train_epochs=10
load_best_model_at_end=True
metric_for_best_model=eval_f1
learning_rate=5e-05
```

The `base` model took approximately 5 minutes, while the `large` model took approximately 12 minutes to train.

For the CES-SD models, the parameters were:

```
dropout=0.3
learning_rate=2e-05
model_name_or_path=albert-xxlarge-v2
num_train_epochs=10
num_warmup_steps=200
per_device_train_batch_size=8
weight_decay=0.005
```

The models took an average of 2 hours to train. The model trained on additional augmented data took slightly longer, around 2.68 hours to train.

For the QA model, to pre-train the model on RECESS, we used the following parameters:

```
model_name_or_path=t5-small
per_device_train_batch_size=12
learning_rate=3e-5
num_train_epochs=5
max_seq_length=512
doc_stride=128
```

For the actual training on SQuAD, we used the following parameters:

```
model_name_or_path=t5-small
per_device_train_batch_size=128
learning_rate=0.001
```

Eval	Model	Overall			Cause			Effect			Signal		
		R	P	F1	R	P	F1	R	P	F1	R	P	F1
Dev	**Baseline**	66.32	59.48	62.71	64.78	66.99	65.87	63.85	64.76	64.30	71.75	48.19	57.65
	+BSS	**71.39**	64.43	67.73	72.60	74.30	73.44	70.67	**73.38**	72.00	**70.70**	47.44	56.78
	+BSS+SC	71.22	**69.81**	**70.51**	72.60	74.30	73.44	70.67	**73.38**	72.00	70.06	**60.44**	**64.90**
	+BSS+SC+DA	70.89	69.25	70.06	**73.35**	**74.71**	**74.02**	**70.97**	73.33	**72.13**	67.31	58.01	62.31
Test	**Baseline**	61.49	61.89	61.69	56.80	65.93	61.03	60.00	62.93	61.43	68.22	57.77	62.56
	+BSS	**67.30**	66.98	67.14	**68.36**	**73.82**	**70.98**	67.58	72.91	**70.14**	**65.80**	55.70	60.33
	+BSS+SC	66.56	**68.86**	**67.69**	**68.36**	**73.82**	**70.98**	**67.73**	72.73	**70.14**	63.21	60.10	61.62
	+BSS+SC+DA	64.43	67.56	65.96	62.79	68.18	65.38	65.75	**73.10**	69.23	64.75	**61.54**	**63.10**

Table 12: Performance Metrics for Cause-Effect-Signal Span Detection. All scores are reported in percentages (%). Highest score per dataset and metric is in bold.

```
num_train_epochs=20
max_seq_length=512
doc_stride=128
```

We change the `model_name_or_path` to the RECESS' pre-trained model's directory for the updated model. For each run to train and predict on SQuAD, it takes around 1.10h.

All experiments were ran on NVIDIA Tesla V100 SXM2 32 GB GPU, CUDA Version: 11.3.