IJCNLP-AACL 2023 @ Bali, Indonesia (November 1 – 4)

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

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoglu, Nelleke Oostdjik, Tommaso Caselli, Tadashi Nomoto, Onur Uca, Farhana Ferdousi Liza, See-Kiong Ng

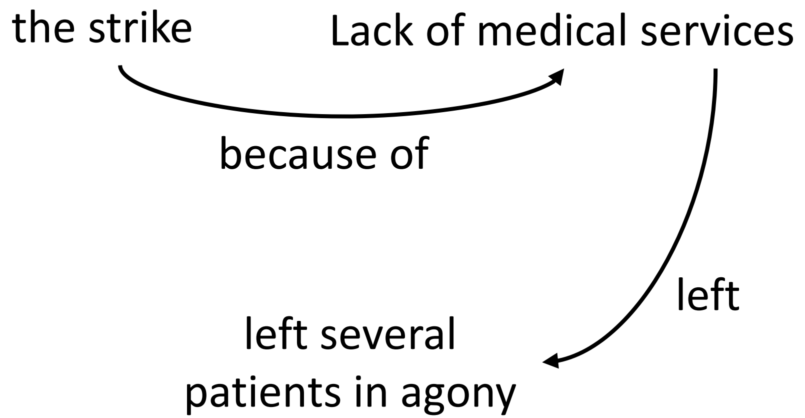
Fiona Anting Tan
Institute of Data Science, National University of Singapore
tan.f@u.nus.edu

Resource for Extracting Cause, Effect, and Signal Spans (RECESS)

- Causal Relations: Semantic relation between where the occurrence of the *Cause* leads to the occurrence of the *Effect*.
- RECESS is larger than other causal text mining benchmarks
  - Our resource comprises of 2,574 causal relations
  - CausalTimeBank (CTB) (Mirza et al., 2014): 318 causal pairs
  - EventStoryLine (ESL) (Caselli and Vossen, 2017):1,770 causal pairs
- We investigated properties of causal relations in text using the rich RECESS annotations.
- Shared task using RECESS was held from May – Sep 2023 to promote research and modelling in this field.

DATASET & ANNOTATION

	Contains *Causal* relations?
Lack of medical services because of the strike left several patients in agony .	✓
KSRTC buses were attacked at ten places .	✗

	Contains *Causal* relations?	**Identified Causal Relations**
Lack of medical services because of the strike left several patients in agony .	✓	

See paper for more details!

Why	Temporal order	Counterfactual	Ontological asymmetry	Linguistic
Not causal if the reader cannot construct a "Why" question based on the Effect.	Not causal if the Cause does not precede the Effect in terms of time.	Not causal if the Effect is equally likely to occur or not occur without the Cause.	Not causal if the reader can swap the Cause and Effect spans.	Likely causal if it can be rephrased into "*X causes Y*" or "*Because of X, Y*".

Figure 1: The Five Tests for Causality (Grivaz, 2010)

Sentence	Causality Tests: Why?	Causality Tests: Temporal Order	Causality Tests: Counter-fact.	Causality Tests: Onto. Asym.	Causality Tests: Linguistic	Label
<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.

RECESS: Data overview

- Expand CNC corpus (Tan et al., 2022)
- News reported from Year 2000 – 2018
- 3,767 sentences in total
- Agreement – Kappa score
    - Binary labels: 34.99
    - Span labels: 42.66
- Disagreements solved by manual curation and discussion

RECESS: Data overview

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

TASK

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

Cause-Effect-Signal Span Detection (CESSD):
Which spans correspond to Cause, *Effect* or *Signal* per causal sentence?

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.

Causal Sentence Classification (CSC)

- **Evaluation Metrics:** Recall (R), Precision (P), Binary F1 (F1), Matthews Correlation Coefficient (MCC)
- **Baseline:** BERT for Sequence Classification (Devlin et al., 2019)
  - bert-base-cased and bert-large-cased
- **Scores:**

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 6: Performance Metrics for CSC.

Cause-Effect-Signal Span Detection (CES-SD)

- **Evaluation Metrics:** Recall (R), Precision (P), Macro F1 (F1)
  - Used FairEval implementation (Ortmann, 2022) of sequence evaluation by word-tokens (Ramshaw and Marcus, 1995) to prevent double penalties of close-to-correct predictions
  - Sentences with multiple causal relations used highest F1 score possible out of all ways to match predicted and true causal relations.

Cause-Effect-Signal Span Detection (CES-SD)

- **Baseline:** Reading comprehension model with BERT-based encoder (Chen et al., 2022)
  - albert-xxlarge-v2
  - Target: $P = [p_{cs}, p_{ce}, p_{es}, p_{ee}, p_{ss}, p_{se}]$
- **Baseline variants:**
  - **Beam-search span selector (BSS)**
  - **Signal Classifier (SC)**
  - **Data augmentation (DA)**

See paper for more details!

Cause-Effect-Signal Span Detection (CES-SD)

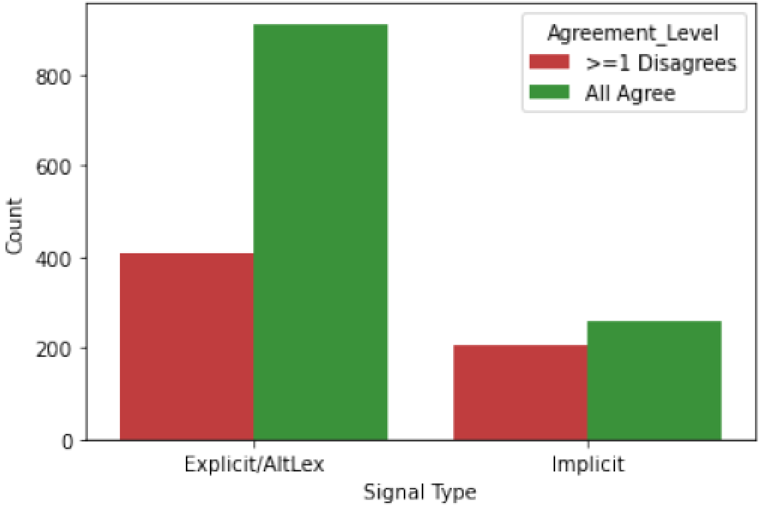
Eval	Model	Overall		
		R	P	F1
Dev	Baseline	66.32	59.48	62.71
	+BSS	**71.39**	64.43	67.73
	+BSS+SC	71.22	**69.81**	**70.51**
	+BSS+SC+DA	70.89	69.25	70.06
Test	Baseline	61.49	61.89	61.69
	+BSS	**67.30**	66.98	67.14
	+BSS+SC	66.56	**68.86**	**67.69**
	+BSS+SC+DA	64.43	67.56	65.96

Table 7: Performance Metrics for CESSD.

When is causality easy/hard to detect?

- Easier if there are causal markers present.

(A) For humans



(B) For model

CSC: Failed to identify *Causal* examples in the proportions

- 8% of Explicit/Altlex
- 17% of Implicit

CES-SD: For 98 perfect predictions,

- 63% were Explicit/Altlex
- 37% were Implicit

Do signals matter?

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 8: End-to-end predictions on example sentences.

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

- RECESS has potential applications for QA, especially for *Why*-Questions.

Templates

- What caused "{effect}"?
- What led to "{effect}“?
- Why did "{effect}" occur?
- What resulted from "{cause}"?
- What happened because of "{cause}"?
- What did "{cause}" cause?

Baseline: t5-small

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 9: QA Performance.

RECESS is a comprehensive corpus annotated for causality at different levels.

- RECESS consists of 3,767 sentences, where 1,982 are causal sentences containing a total of 2,754 causal relations.
- Our annotation guidelines cover a broad range of linguistic, semantic, and syntactic structures for causal relations.
- 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
- We also performed investigations of causal relations in text.

Thank you.

- Link to repository: https://github.com/tanfiona/CausalNewsCorpus
- Please share your feedback with us:
  Fiona Anting Tan (tan.f@u.nus.edu)