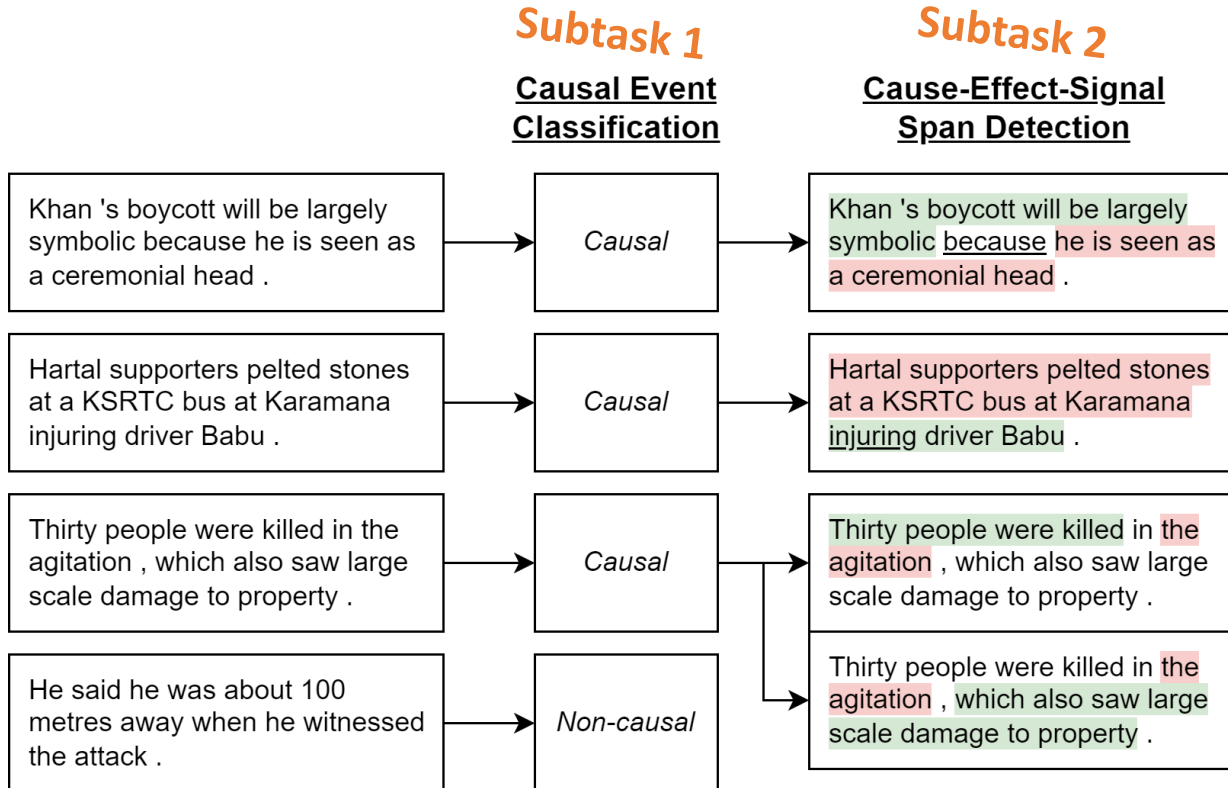# Event Causality Identification with Causal News Corpus - Shared Task 3, CASE 2023

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoglu, Nelleke Oostdijk, Onur Uca, Surendrabikram Thapa, and Farhana Ferdousi Liza.

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

01 Event Causality Identification Shared Task involved two subtasks related to Classification and Span Detection.

Subtask 1
Causal Event Classification

Subtask 2
Cause-Effect-Signal Span Detection

Khan 's boycott will be largely symbolic because he is seen as a ceremonial head .	*Causal*	Khan 's boycott will be largely symbolic because he is seen as a ceremonial head .
Hartal supporters pelted stones at a KSRTC bus at Karamana injuring driver Babu .	*Causal*	Hartal supporters pelted stones at a KSRTC bus at Karamana injuring driver Babu .
Thirty people were killed in the agitation , which also saw large scale damage to property .	*Causal*	Thirty people were killed in the agitation , which also saw large scale damage to property .
He said he was about 100 metres away when he witnessed the attack .	*Non-causal*	Thirty people were killed in the agitation , which also saw large scale damage to property .

02

Subtask 1 worked directly on the Causal News Corpus V2 (CNC-V2) / RECESS (Tan et al., 2023).

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

Subtask 1 Data Summary Statistics.

- Data Source:
  Causal News Corpus V2
  a.k.a. RECESS (Tan et al., 2023)

- A sentence is *Causal* if "one argument provides the reason, explanation or justification for the situation described by the other"(Webber et al., 2019) and contains at least a pair of events.

03 Subtask 2 worked directly on the Causal News Corpus V2 (CNC-V2) / RECESS (Tan et al., 2023).

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

Subtask 2 Data Summary Statistics.

- A **Cause** is a reason, explanation or justification that led to an **Effect**.

- **Signals** are words that help to identify the structure of the discourse.

04

We provided multiple evaluation metrics, but model performance was eventually ranked by F1.
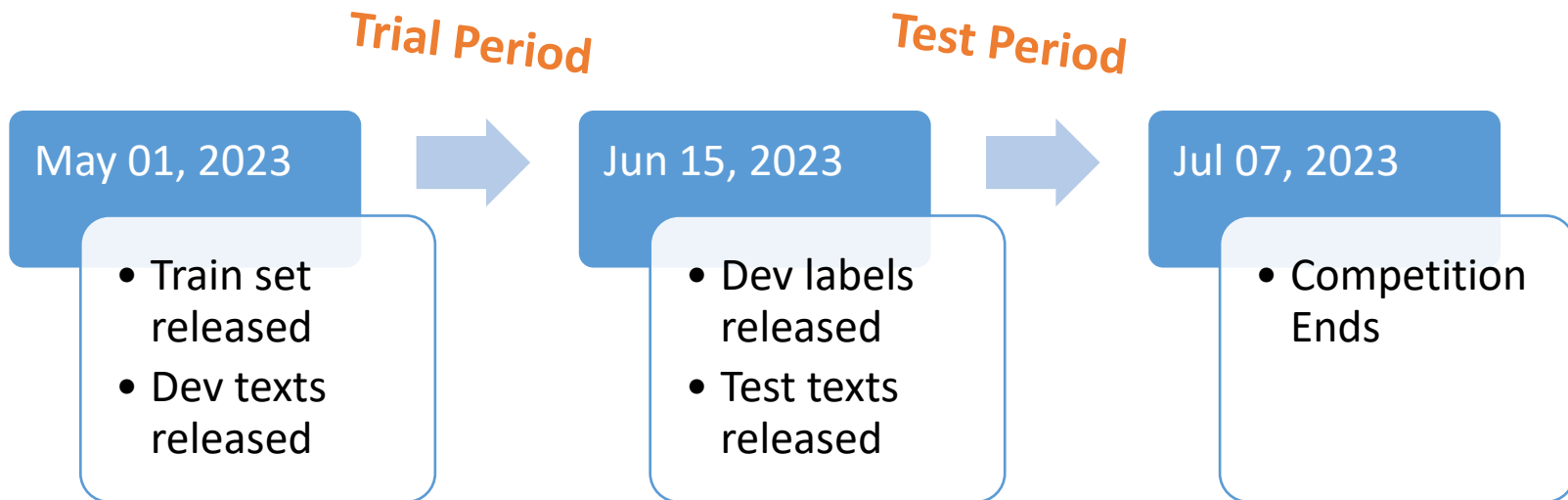
- The following evaluation metrics were provided:
  - Subtask 1: Accuracy, Binary Precision (P), Binary Recall (R), Binary F1 and Matthews Correlation Coefficient
  - Subtask 2: Macro P, R and F1 based on word labels

- Leader board was ranked by F1 for both tasks

- For Subtask 2, to handle predictions for examples with multiple causal relations:
  - If more predictions (p) are provided than true relations (n), we only consider the first n relations.
  - If fewer predictions (p) are provided than true relations (n), we assume the missing n-p relations have all "Other" tokens.
  - Once n=p, we calculate every combination of pairs of prediction and true relations and retain the combination that gives us the highest score.

05

We used the Codalab website to host our competition.

06 **Timeline**

Trial Period

Test Period

May 01, 2023

- Train set released
- Dev texts released

Jun 15, 2023

- Dev labels released
- Test texts released

Jul 07, 2023

- Competition Ends

Timeline of competition.

07 **There were 23 registered participants throughout the competition.**

29 applied	23 registered	10 participated	6 submitted papers	4 accepted papers

Number of teams per stage of competition.

08 The best F1 score for Subtask 1 was 84.66%.

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

Subtask 1 Leaderboard.

10 **The best F1 score for Subtask 2 was 72.79%.**

Rank	Team Name	Codalab Username	Overall		
			R	P	F1
1	BoschAI (Schrader et al., 2023)	timos	**63.98**	**84.42**	**72.79**
2	1Cademy Baseline	tanfiona	59.18	60.25	59.71
3	CSECU-DSG (Hossain et al., 2023)	csecudsg	36.12	40.00	37.96
4	-	pakapro	0.00	0.00	0.00

Subtask 2 Leaderboard.

Conclusion & Future Work

- Two subtasks:
    1) Causal Event Classification, and
    2) Cause-Effect-Signal Span Detection.

- Each subtask attracted predictions from models that beat our baselines.

Thank you.

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