Shrome at Touché: Soft-Vote Ensembling and Counter-Causal Augmentation for Causality Extraction

Shayan Sooratgar & Roham Zendehdel Nobari

Department of Computational Linguistics, University of Zurich · shayan.sooratgar@uzh.ch · roham.zendehdelnobari@uzh.ch

Code, prompts & generated data
github.com/rzninvo/NegaCausal

Held-out test results

Countercausal News Corpus (CCNC), official TIRA evaluation

Extraction · granularity-adj. F1*
0.728
highest of all submissions · baseline 0.665 · other team 0.587

Polarity · macro F1
0.817
best team · baseline 0.839 · other team 0.778

Detection · F1
0.869
highest recall (0.890) · others 0.884, 0.872

1 Causal words do not always mean causation

procausal
The strike caused massive delays.

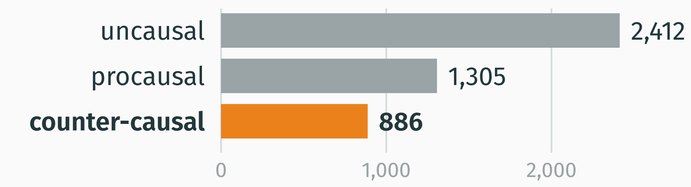
counter-causal
It is falsely believed that the strike caused massive delays.

Same cause, same effect, same verb. A model that trusts surface cues such as *caused* or *led to* accepts both as causal and gets the polarity of the second wrong.

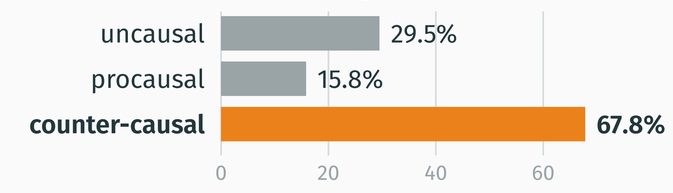
2 Three subtasks, one rare class

Subtask	Output	Train	Dev
Detection	causal / uncausal	3,075	340
Extraction	cause & effect spans	3,075	340
Polarity	pro / counter / uncausal	4,603	502

Polarity training rows



Dev sentences with explicit negation cues



3 One model per subtask, one rule

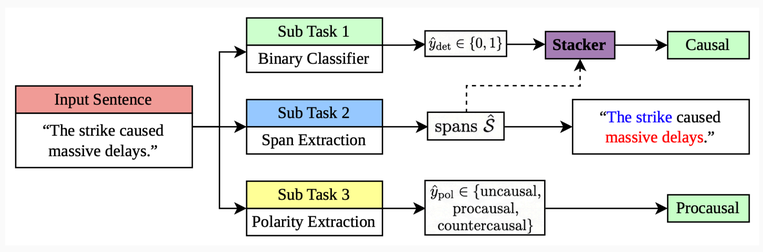
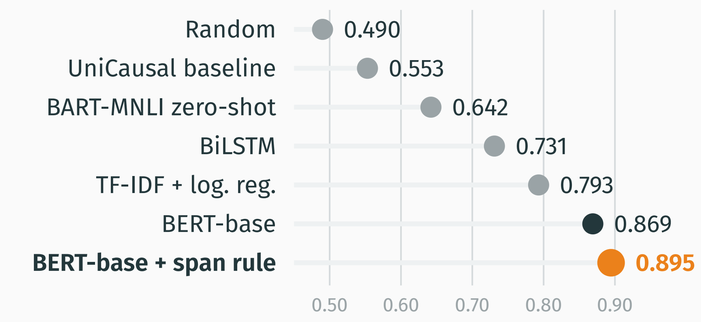


Figure 1 of the paper.

Stacker rule: predict *causal* only if the detector says causal **and** extraction finds ≥ 1 span. Models are otherwise trained independently.

4 Checking for spans adds 2.6 points

Detection F1 on dev (causal class)



31 false positives fixed | **15 new FN**

The rule flips 46 dev sentences to uncausal.

5 Three BILOU layers on RoBERTa-large

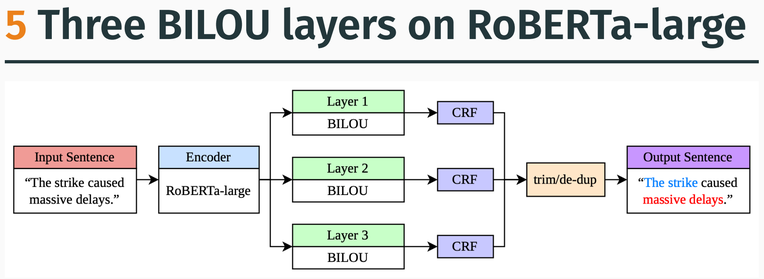


Figure 3 of the paper: one extraction model. Gold spans are coloured onto ≤ 3 layers, so overlapping spans fit.

Best single-model exact-match F1 on dev

≈ 11,500 training rows after augmentation (TAPT seed ≈ 13,400 with UniCausal). CRF parameters use a 6× higher learning rate.

A Contribution 1 · Average the scores, then decode once

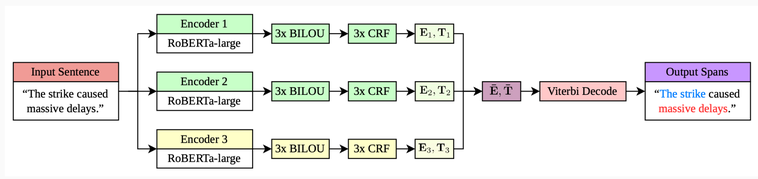


Figure 4 of the paper: emission-level soft-vote over three seeds.

Hard vote

1. each seed decodes its own spans
2. majority vote over spans
3. near-misses are discarded

Soft vote (ours)

1. average emissions, transitions, start/end scores
2. one Viterbi decode per layer
3. near-misses survive

Three seeds score a span start at 0.80, 0.60 and 0.70: the averaged decoder sees **0.70**, even if no single seed emitted that start.

Stage (dev)	Exact F1	Overlap F1
Single seed (123)	0.5742	0.7692
Hard span vote, matched 3 seeds	0.5781	0.7634
Soft vote, same 3 seeds	0.5888	0.7730
Soft vote, production panel	**0.5933**	**0.7800**
+ NLI span verifier (within noise)	0.5953	0.7781

Same models, only the combiner changes: **+1.07 pp exact F1**.

Ensemble beats every member

Seed	Recipe	Exact	Overlap
123	base	0.5742	0.7692
2026	base	0.5505	0.7569
42	TAPT + UniCausal + boundary smoothing	0.5260	0.7584
Soft vote of all three		**0.5933**	**0.7800**

TAPT helped only seed 42 and hurt three other seeds; it is kept for encoder diversity, not as an endorsement.

B Contribution 2 · Teach the model how causation is denied

direct negation (h)	**negated** context (i)	**lack of** counterfactuality (b)
lack of effect (c)	**implicit** lack of effect (c)	**negative** causation (f)
usual inverse effect (e)	**coincidence** (b)	**temporal** order (a)

Nine patterns adapted from the categories of Hagen et al. (letters). (f) is kept on purpose as a hard surface-form confusable.

CCNC train sentence → **GPT-5.4, one pattern** → **structural check** → **polarity training**

Check: cause/effect span text byte-identical, exactly one <e0> and one <e1> pair. It checks structure, not meaning.

93.2%
accepted (2,433 / 2,610)

886 → 3,319
counter-causal training rows

.766 → .890
dev macro F1, seed 42

Endpoints use different checkpoint criteria; with matched selection the second pass adds +2.4 pp.

6 Typed markers, four-seed ensemble

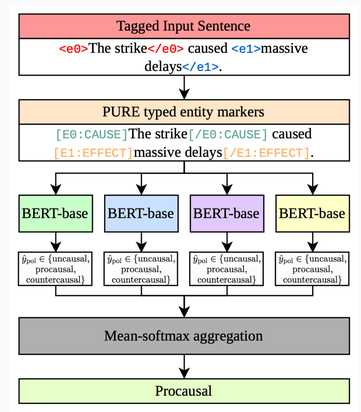
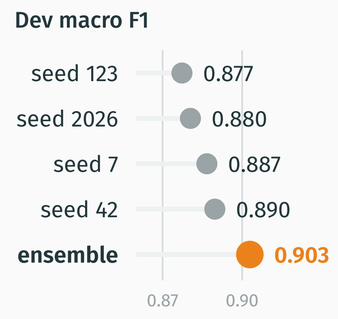


Figure 5 of the paper.

Dev macro F1



gold ↓ pred →	un	pro	cnt
uncausal	**243**	12	9
procausal	8	**106**	6
counter	8	5	**105**

Seed 42, dev. Main error: uncausal predicted causal (21).

7 Test set: dev was optimistic

Subtask	Metric	Dev	Test	Δ
Detection	F1	0.895	0.869	−2.6
Extraction	exact F1 (TIRA)	0.593	0.473	−12.0
Polarity	macro F1	0.903	0.817	−8.6

Panel, rule, thresholds and ensemble size were all selected on dev.

Takeaways

1. **Average CRF scores before you decode.**
2. **Teach the model the ways causation is denied.**
3. **Check dev gains on test.**

Next: a negation-scope feature for polarity; new encoders and span architectures for extraction.

*Organizers' final causal-only evaluation · CCNC: Hagen et al., 2025