Parsing Causal Relations in Social Science Publications

Toward Transparent, Reproducible Meta-Science

Rasoul Norouzi, Bennett Kleinberg, Jeroen Vermunt, & Caspar Van Lissa

Tilburg University

github.com/rasoulnorouzi/JointLearning

Why Meta-Science Needs to Care About Causality

The Foundation

Causality is the foundation of theory and intervention. It’s how we explain *why* things happen.

The Problem

Causal claims are often hidden in ambiguous text: "might lead to," "is associated with."

The Consequence

Without causal clarity, we have no cumulative science. We can't systematically map or test our theories.

From Statistical Effects to Causal Understanding

> *"The causal revolution lets us ask* ***why****, not just what."*
>
> — Judea Pearl, 2018

Meta-Science Today

Focuses heavily on synthesizing **statistical results**:

- Averaging effect sizes
- Testing replicability
- Detecting publication bias

The Missing Piece

We rarely synthesize the **causal claims** that studies propose.

- What theories are being tested?
- Which causal mechanisms are proposed?
- Where do theories agree or conflict?

Bridging this gap is the key to cumulative theory.

Manual Causal Coding: Essential but Unsustainable

Scale

The literature grows faster than humans can read. Comprehensive review is impossible.

Bias

Humans unconsciously confirm familiar theories (confirmation bias) and overlook alternatives.

Inconsistency

Ambiguous language means even trained coders disagree ~20% of the time. The process is a black box.

Cost

Manual review is incredibly slow and expensive, diverting expert resources from new research.

We need a system that scales human reasoning without losing interpretability.

The Case for Auditable Automation

The goal is to replicate expert annotators.

We need a system that can:

Scale Up
Map causal claims from thousands of papers.

Preserve Transparency
Trace every extracted claim back to its source text.

Enable Replication
Allow any researcher to inspect, verify, and build upon the results.

Innovation 1: Better Data & Smarter Schema

Domain-Specific Dataset

3,014 sentences annotated from social science papers to teach the model the nuances of our field's language.

Expanded Schema for Causal Chains

Our new `CE` tag identifies mediating variables.

"Drought led to crop failure, which caused unrest."

Drought (C) → crop failure (CE) → unrest (E)

Innovation 2: A Bi-Directional, Self-Correcting Architecture

Previous: Sequential Models

Tasks run one-by-one. An error in an early step cascades and cannot be corrected.

1. Identify Sentence

✕ ERROR ↓

2. Extract Spans

↓

3. Link Relations

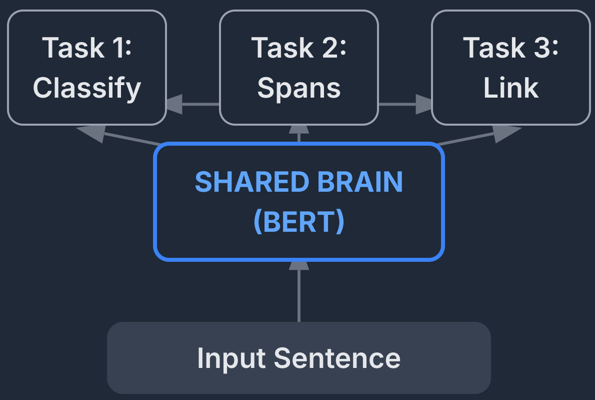
Example Sentence:

"This study reviewed 120 experiments on social ostracism."

Result: Wrongly classified as Causal. Error propagates.

Our Model: A Bi-Directional Architecture

A shared "brain" processes the text, and all tasks communicate to refine the final decision.



Same Sentence:

"This study reviewed 120 experiments on social ostracism."

Result: Task 2 & 3 find no pairs, telling Task 1 to self-correct. Final output is correct.

The Model Achieves Human-Level Reliability

Task	Human–Human α	Model–Human α
Causal Detection	0.65	0.63
Cause Spans	0.83	0.77
Effect Spans	0.91	0.90
Linking	0.83	**0.96**
Overall Agreement	0.80	**0.81**

The model's consistency is statistically indistinguishable from trained human experts. It's not just automating work—it's codifying expertise in a reproducible way.

It Doesn’t Just Match Scores — It Imitates Judgment

The model learned the same patterns of difficulty and ambiguity that humans face.

Harder to Identify: Causes

Cause spans are often longer, more abstract, and grammatically complex. Humans show less agreement here.

Human-Human α: 0.83

Model-Human α: 0.77

Easier to Identify: Effects

Effect spans are typically more concise, concrete, and appear after causal verbs. Humans are more consistent.

Human-Human α: 0.91

Model-Human α: 0.90

The model learned to hesitate where humans hesitate. This gives us a computational mirror of our own decision-making process.

A Robust & Well-Calibrated Tool for Science

Excellent Overall Performance

F1 = 0.78

The model demonstrates high precision and recall across all tasks, making it a reliable tool for evidence synthesis.

Performance is also extremely stable across different confidence thresholds, indicating a well-calibrated and trustworthy model.

From Extraction to Evidence Mapping

Each output is a traceable unit of evidence.


```json
{
  "text": "exercise improves physical health and mental well-being",
  "causal": true,
  "relations": [
    {
      "cause": "exercise",
      "effect": "physical health"
    },
    {
      "cause": "exercise",
      "effect": "mental well-being"
    }
  ]
}
```

An Open & Verifiable Process

Building Trust Through Transparency

- **Open Dataset & Code:** Allows for full replication and inspection of our work.
- **Published Schema:** The rules for annotation are explicit and can be debated or extended.
- **Documented Methods:** All training and validation procedures are described for independent verification.

This isn't a "black box"—it's a glass box designed for scientific scrutiny.

Future Directions: The Road Ahead

Ecosystem & Validation Tools

- **Interactive Review Platform:** Develop tools that integrate automated extraction with expert oversight.
- **Active Learning:** Use the model to prioritize ambiguous cases for human review, improving annotation quality.
- **Robustness Testing:** Evaluate the model across diverse social-science subfields to ensure generalizability.

How Meta-Scientists Can Join

This isn’t just a tool — it’s infrastructure for the community.

Contribute

Contribute annotated data from your domain to improve model robustness and help refine the annotation schema.

Validate

Test the model on your own corpus to evaluate its performance and identify areas for improvement.

Build

Connect the structured outputs to your meta-analytic pipelines or build new dashboards for evidence synthesis.

The Meta-Scientific Payoff

Structured causal data enables a new class of meta-science.

- Map theory evolution over time.
- Identify under-studied or contradictory claims as replication targets.
- Detect citation bias where some causal paths are reinforced but others are ignored.
- Build Causal Question-Answering systems to synthesize evidence on demand.

Thank You & Questions

Scan for Code & Data

github.com/rasoulnorouzi/JointLearning

Ask me about:

Reliability, model thresholds, annotation challenges, & future collaborations.