Conversational AI, Assignment-01, 2023AC05475,

Shailesh Singh, 06/06/2025.

→ Preliminary content:

1. BITS Student ID - 2023AC05475
2. Name - Shailesh Kumar Singh
3. Group No. - 43
4. Title of the Research Paper & Authors: Causal Inference for Human-Language Model Collaboration.
5. Online link: https://arxiv.org/pdf/2401.00207

→ Report - Core Analysis of the Research paper:

1. Problem Resolved & key findings:

The paper addresses the challenge of measuring the impact of human edits in human-language model (LM) collaboration. When people edit LM-generated text, its unclear how much they actully improve or change the outcome.

To solve this the authors propose a causal framework and introduce a new metric:

Key finding:- human contributions can be measured & interpreted more accurately using this method.

To solve this the authors propose a new causal framework to measure the true influence of human edits using a metric called Incremental Stylistic Effect (ISE). This metric captures how small stylistic changes by human causally affect the final content quality.

They also develop a novel system called CausalCollab which uses counterfactual simulations in a differentiable style space to estimate ISE. Experiments show this method outperforms traditional baselines in detecting meaningful human contributs. The framework makes human input interpretable and measurable which is a break through in collaborative NLP systems.

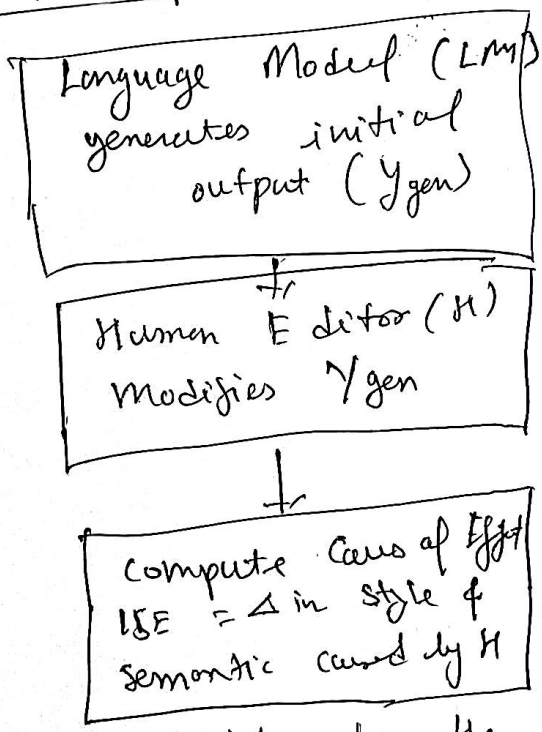
2). Methodology & Architecture Summary

The authors use causal interfence theory to understand human contributions in collaborative writing. Instead of treating edits as simple binary actions. they map human edits into a continuous vector space called a style space. Each edit is treated as a treatment and the final output is the outcome in a causal graph.

To measure the effect they use a differential prediction model that accepts the style vector & produces the output. By applying small perturbations to the style vector, the system simulates counterfactuals. What would have happened if the edit was different. They, using causal gradients, they compute the incremental stylistic which reflects how influential the edit was on the final outcome. Their proposed system, causal collab, consist of:

- A style encoder that converts human edits into vectors.
- A prediction model that produces outcomes.
- And a causal estimates that computes ISE by comparing factorial & counterfactual outputs.

3 Visual Component.



- Y_{gen} = Text generated by the model
- Y_h = Final text after human edits.
- ISE = Quantifies how much the human's changes influenced the final output.
- Causal assumption: Human edits are an intervention on LM output
- Goal understudy who contributed what in human AI

page (03)

- Perturbations are applied → Counterfactual generator → predictive model → ISE Estimator.
- Boxes represent components and arrows show the data flow through the system.

These diagrams explain how human edits are encoded perturbed and evaluated for their effect on the final text using counterfactual reasoning.

4 Critical Evaluation

This paper is innovative and timely. One major strength is its ability to provide interpretable and quantitative insights into human contributions in AI-assisted writing. Traditional systems only look at surface-level edits but this work goes deeper by measuring how much those edits really matter. The ISE surface level edits, but this work goes deeper by measuring how much those edits really matter.

Another strength is the differentiable design, which means the system can be trained and integrated into modern NLP architecture. The experimental results although on controlled datasets show clear improvements in understanding human impact.

However, there are weakness & limitations. The biggest concern is generalizability. the current

System is evaluated in simplified or controlled environments. Its unclear how well it would perform in real-world, large-scale editing scenarios with noisy or domain-specific context - Also the method assumes that human edits can be encoded in a low-dimensional continuous space, which might not always hold true - espicially for complex structural or semantic changes.

The system also assumes that the only significant influence on the final outcome is style, whereas content based or logical changes might also play a major role but are not directly modeled - Another assumption is that perturbing the style vector linearly will simulate plausible.

This paper has been revised by all the group mates and specially by such individuals from group A3.