Counterfactual Data Augmentation for Model Robustness

Johnny Ma, Nitish Joshi, Sam Bowman, He He

March 5th, 2021

Experimental Setup

- Fine-tune BART on MNLI conditional pairs (premise + hypo, mask hypo) MLM task for 3 epochs.
- Apply one of three masking strategies
  a. Data Slice (prepositions)
  b. Gradient based (RoBERTa l1-norm w.r.t. gold label)
  c. Content Words (POS Nouns, Verbs, Adjs)
- Generate over all 3 strategies for 48-hours.
- Label with RoBERTa fine-tuned on MNLI.
- Designate flip **x** certainty (< 80%) classes.

	Generations	Premises
Content Words	66,652	3,906
Data Slices	53,236	8,272
Gradient	66,027	6,325

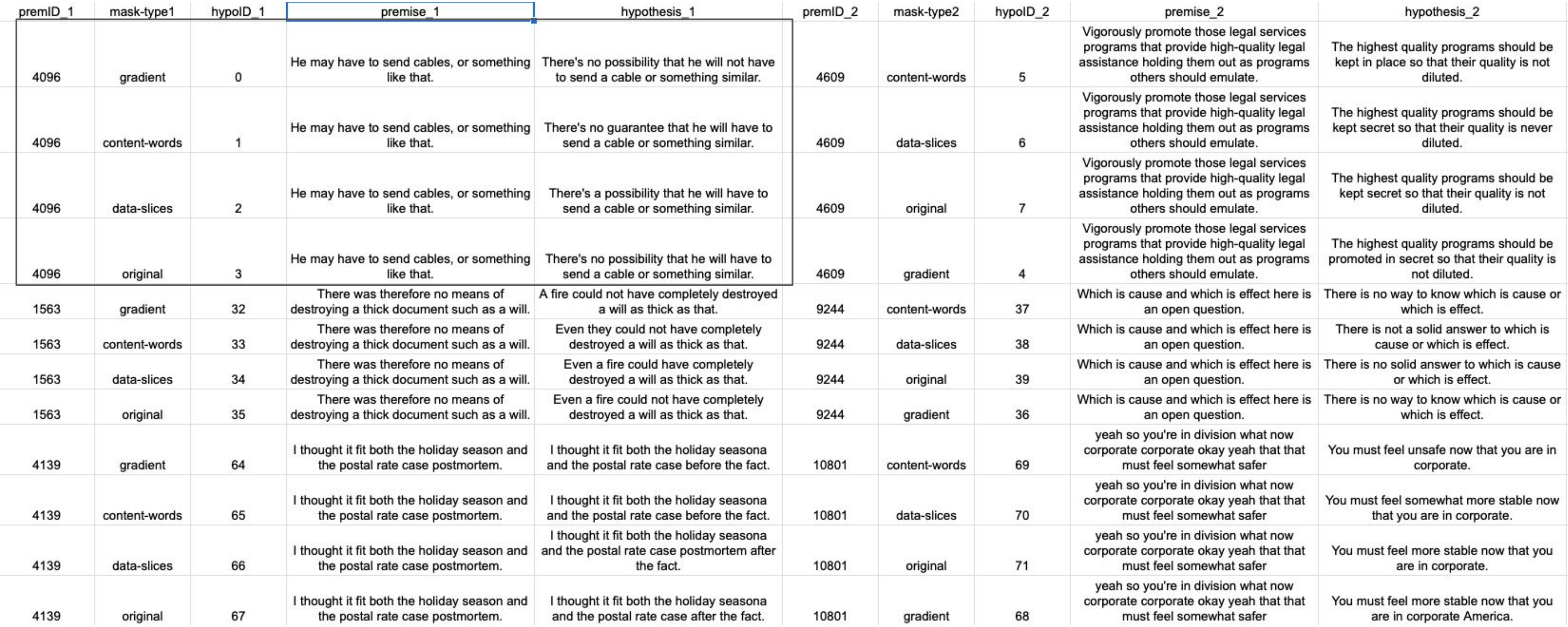
Table 1: Generations by Masking Strategy, 48 hours.

- Pick premises that have **at least 3** flip-classes **across 3 masking strategies.**
  - Obtains ~189 shared premises
  - Only ~85 shared premises with all **4 classes**
- Sample 3:1 uncertain vs. certain
  - Produces slightly unbalanced classes, but rough 3:1 ratio.
- Include original premise + hypothesis in data for validation.
- Transform 4 * 189 premises into 92 HITs with 8 premise-hypos each.
- HIT 1: [P1/M1, P2/M2, P3/M3, P4/O, ...P8/MO]
- HIT 2: [P1/M2, P2/M3, P3/O, P4/M1, ... P8/M1]

Counts	Certain-Same	Certain-Flip	Uncertain-Same	Uncertain-Flip
Content Words	35	37	54	63
Data Slices	51	41	50	47
Gradient	28	29	54	78
Original	**189**			

Table 2: Counts by Flip-class and Masking Strategy, Data in Trial

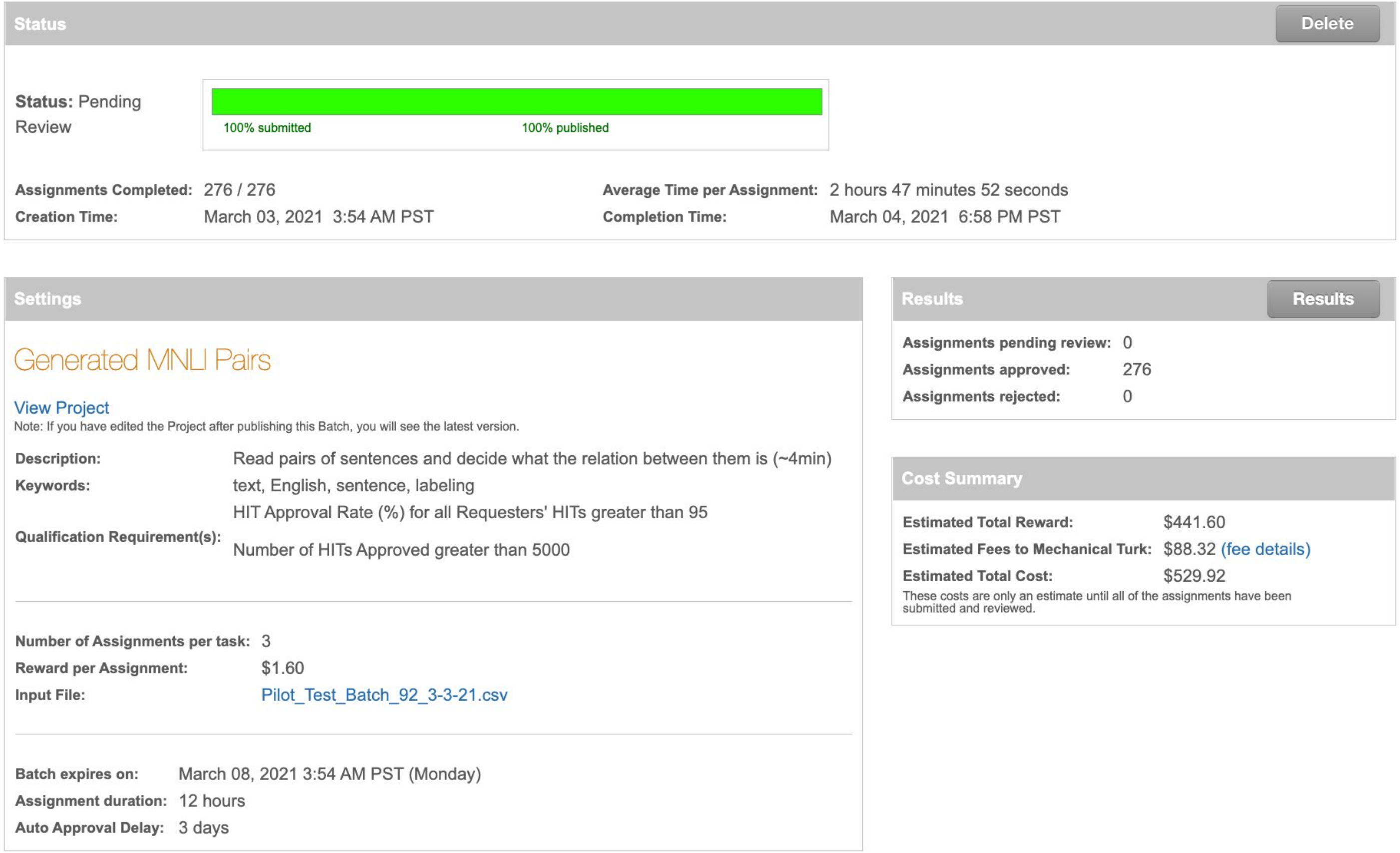
https://docs.google.com/spreadsheets/d/1hZJ1eLQdTi-bXtqQOoEzbdj_7CCG9X0taNS_ir81W1w/edit?usp=sharing



premID_1	mask-type1	hypoID_1	premise_1	hypothesis_1	premID_2	mask-type2	hypoID_2	premise_2	hypothesis_2
4096	gradient	0	He may have to send cables, or something like that.	There's no possibility that he will not have to send a cable or something similar.	4609	content-words	5	Vigorously promote those legal services programs that provide high-quality legal assistance holding them out as programs others should emulate.	The highest quality programs should be kept in place so that their quality is not diluted.
4096	content-words	1	He may have to send cables, or something like that.	There's no guarantee that he will have to send a cable or something similar.	4609	data-slices	6	Vigorously promote those legal services programs that provide high-quality legal assistance holding them out as programs others should emulate.	The highest quality programs should be kept secret so that their quality is never diluted.
4096	data-slices	2	He may have to send cables, or something like that.	There's a possibility that he will have to send a cable or something similar.	4609	original	7	Vigorously promote those legal services programs that provide high-quality legal assistance holding them out as programs others should emulate.	The highest quality programs should be kept secret so that their quality is not diluted.
4096	original	3	He may have to send cables, or something like that.	There's no possibility that he will have to send a cable or something similar.	4609	gradient	4	Vigorously promote those legal services programs that provide high-quality legal assistance holding them out as programs others should emulate.	The highest quality programs should be promoted in secret so that their quality is not diluted.
1563	gradient	32	There was therefore no means of destroying a thick document such as a will.	A fire could not have completely destroyed a will as thick as that.	9244	content-words	37	Which is cause and which is effect here is an open question.	There is no way to know which is cause or which is effect.
1563	content-words	33	There was therefore no means of destroying a thick document such as a will.	Even they could not have completely destroyed a will as thick as that.	9244	data-slices	38	Which is cause and which is effect here is an open question.	There is not a solid answer to which is cause or which is effect.
1563	data-slices	34	There was therefore no means of destroying a thick document such as a will.	Even a fire could have completely destroyed a will as thick as that.	9244	original	39	Which is cause and which is effect here is an open question.	There is no solid answer to which is cause or which is effect.
1563	original	35	There was therefore no means of destroying a thick document such as a will.	Even a fire could not have completely destroyed a will as thick as that.	9244	gradient	36	Which is cause and which is effect here is an open question.	There is no way to know which is cause or which is effect.
4139	gradient	64	I thought it fit both the holiday season and the postal rate case postmortem.	I thought it fit both the holiday seasona and the postal rate case before the fact.	10801	content-words	69	yeah so you're in division what now corporate corporate okay yeah that that must feel somewhat safer	You must feel unsafe now that you are in corporate.
4139	content-words	65	I thought it fit both the holiday season and the postal rate case postmortem.	I thought it fit both the holiday seasona and the postal rate case before the fact.	10801	data-slices	70	yeah so you're in division what now corporate corporate okay yeah that that must feel somewhat safer	You must feel somewhat more stable now that you are in corporate.
4139	data-slices	66	I thought it fit both the holiday season and the postal rate case postmortem.	I thought it fit both the holiday seasona and the postal rate case postmortem after the fact.	10801	original	71	yeah so you're in division what now corporate corporate okay yeah that that must feel somewhat safer	You must feel more stable now that you are in corporate.
4139	original	67	I thought it fit both the holiday season and the postal rate case postmortem.	I thought it fit both the holiday seasona and the postal rate case after the fact.	10801	gradient	68	yeah so you're in division what now corporate corporate okay yeah that that must feel somewhat safer	You must feel more stable now that you are in corporate America.

Figure 1: A few rows of HITs

https://requester.mturk.com/batches/4352801



Status

Status: Pending Review

Assignments Completed: 276 / 276

Creation Time: March 03, 2021 3:54 AM PST

Average Time per Assignment: 2 hours 47 minutes 52 seconds

Completion Time: March 04, 2021 6:58 PM PST

Settings

Generated MNLI Pairs

View Project

Note: If you have edited the Project after publishing this Batch, you will see the latest version.

Description: Read pairs of sentences and decide what the relation between them is (~4min)

Keywords: text, English, sentence, labeling

Qualification Requirement(s): HIT Approval Rate (%) for all Requesters' HITs greater than 95
Number of HITs Approved greater than 5000

Number of Assignments per task: 3

Reward per Assignment: $1.60

Input File: Pilot_Test_Batch_92_3-3-21.csv

Batch expires on: March 08, 2021 3:54 AM PST (Monday)

Assignment duration: 12 hours

Auto Approval Delay: 3 days

Results

Assignments pending review: 0

Assignments approved: 276

Assignments rejected: 0

Cost Summary

Estimated Total Reward: $441.60

Estimated Fees to Mechanical Turk: $88.32 (fee details)

Estimated Total Cost: $529.92

These costs are only an estimate until all of the assignments have been submitted and reviewed.

Figure 2: MTurk Summary Page

MTurk Processing and Purifying

- 92 HITs, 184 premises * (3+1) masking strategies (original + 1)
- Assign majority label when **not 3** unique labels per premise-hypothesis pair.

- Remove premises where **even 1** generated hypothesis has no majority.
  - (184 -> 112)
- Merge with labeled data using unique hypothesis ID.

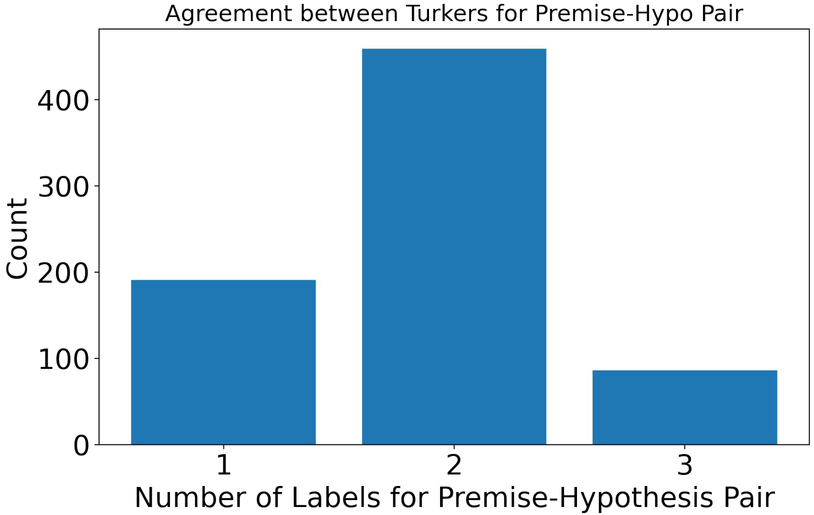


Figure 4: Annotator Disagreement

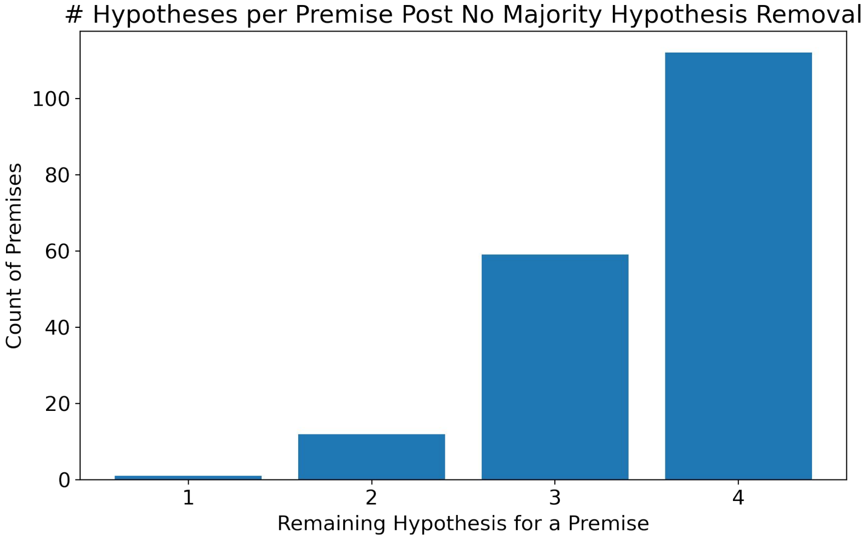


Figure 5: Premises Affected by Disagreement

Results: Purest

Model Wrong %	Certain-Same	Certain-Flip	Uncertain-Same	Uncertain-Flip
Content Words	30%	39%	60%	58%
Data Slices	26%	25%	52%	58%
Gradient	37%	43%	38%	46%
Original	**27% (annotator wrong vs. gold)**			

Results: Unpure

Model Wrong %	Certain-Same	Certain-Flip	Uncertain-Same	Uncertain-Flip
Content Words	35%	37%	56%	51%
Data Slices	32%	30%	44%	56%
Gradient	41%	44%	46%	49%
Original	**27% (annotator wrong vs. gold)**			

Results Discussion

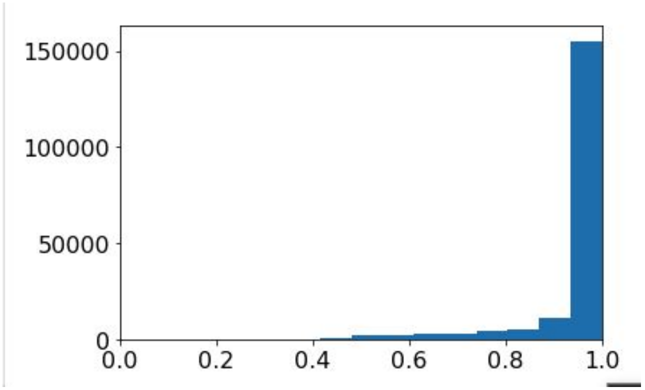
- Large difference between model uncertain and model certain
- Minimal differences between masking strategies
  - Data slices seems best, gradient is not great (likely better than MiCE random)
- Data slices heavily limit ability to pick premises across all 3 strategies, could be picking model initially uncertain data instances.
- Hard to make conclusions given annotator noise.

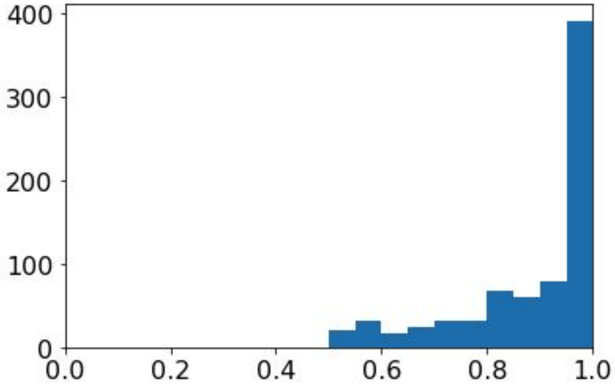
Lessons from Pilot

- **Seems like annotator noise on MNLI is rather high.**
  - Increase #Turkers per HIT 3 -> 5, keep majority vote
  - Stricter incoming qualifications
  - Set up a training/qualification round
  - Filter for high quality workers based on “gold” original examples.
- Run test on model initially uncertain data instances?
  - Distribution of flip-classes are different, maybe quality of examples are as well.
- Compare various sampling strategies next?
- Add measure of fluency?
  - seems like most generations are relatively good at this stage.

Why are we seeing low annotator quality?

1. **Premises chosen are difficult.**
2. Hypotheses generated are difficult.
3. Workers are bad at the task.
4. The task is inherently noisy and requires annotator agreement.
5. We don’t have a filter for high quality submissions.


150000
100000
50000
0
0.0
0.2
0.4
0.6
0.8
1.0


400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0