

APPENDIX

A. Negation examples

Appendix Table I shows one randomly sampled example per available negation method when applied onto the CSci corpus. As shown, most examples fall into ‘VB_3.1’, ‘VB_5.1’, ‘JJ_1.3’ and ‘VB_1.2’ types, for which the templates in Algorithm 1 work well for¹⁴ For rarer method types, like ‘VB_2.1’, the templates seem to work poorly. Further investigation shows that the error arose from the POS tagging step: “Both” was tagged as a VB but should have been a DT or CC, for which, we have no template for at the moment, so the example would have been correctly skipped. As for ‘VB_4.1’, the negated example is unnatural but not grammatically wrong.

B. De-duplication

After appending original sentences with edits, we conduct de-duplication. Appendix Table III shows problematic duplicates that had differing labels. The original CSci corpus contained 7 duplicate sentences instances which were removed. 6 of them were exact duplicates (same label, same sentence), while the last 1 (sentence S/N 1) was duplicated with different labels (c_0 and c_2). We manually changed this to keep only the c_0 label. The total data size thus reduces from $n=3061$ to $n=3054$. We also take this chance to highlight some concerns that some sentences in CSci were labelled contrary to how we understood them.

Subsequent duplicates were handled via rule-based removal. The motivation was to ensure identical sentences do not have different labels which adds noise to our training. Our assumption is that if an edit was performed but remained identical to the original, the original must have been mislabelled sentence. We note that our rule-based de-duplication cannot accommodate multi-label cases, as there was one sentence (S/N 4) that correctly reflected both c_0 and c_1 labels in different parts of the sentence, but due to de-duplication, we only kept the c_0 label.

C. Additional figures and tables

¹⁴We highlight the main POS tags used and mentioned: VB (verbs, e.g. ‘eating’), JJ (adjective, e.g. ‘big’), IN (preposition or subordinating conjunction, e.g. ‘by’), DT (determiner, e.g. ‘he’), CC (coordinating conjunction, e.g. ‘and’), MD (modal, ‘should’).

Algorithm 1: NegationRules – Causal negation scheme

Input: *edit_id, text_ids, text, pos, sentid2tid, max_try=2, curr_try=0*

Output: *text, method, edit_id*

```
1 curr_try ← curr_try + 1
2 curr_pos, curr_word ← pos[edit_id], text[edit_id]
3 prev_pos, prev_word ← pos[edit_id - 1], text[edit_id - 1] if valid else None
4 next_pos, next_word ← pos[edit_id + 1], text[edit_id + 1] if valid else None
5 while curr_try ≤ max_try do
6   if curr_pos = VB then
7     if curr_word = AuxilliaryType then
8       if edit_id = max(text_ids) then
9         Insert *not* in front of curr_word // Method 'VB_1.1'
10      else if next_word = DeterminerType then
11        Replace next_word with *no* // Method 'VB_1.2'
12        edit_id ← edit_id + 1
13      else if next_word = NounType then
14        Insert *not* behind of curr_word // Method 'VB_1.3'
15      else if next_pos = VB then
16        Insert *no* behind of curr_word // Method 'VB_1.4'
17      else if edit_id = min(text_ids) then
18        Replace curr_word with *Not* + lowercased curr_word // Method 'VB_2.1'
19      else if prev_word = NounType then
20        Replace curr_word with *did not* + lemma(curr_word) // Method 'VB_3.1'
21      else if edit_id = max(text_ids) then
22        Insert *not* in front of curr_word // Method 'VB_4.1'
23      else if prev_word = AuxilliaryType next_pos = IN|TO then
24        Insert *not* in front of curr_word // Method 'VB_5.1'
25    else if curr_pos = NN then
26      Get head_id of head word of curr_word based on dependency tree text, method, edit_id ←
        NegationRules(head_id, text_ids, text, pos, sentid2tid, curr_try)
27    else if curr_pos = JJ then
28      if edit_id = max(text_ids) then
29        Insert *not* in front of curr_word // Method 'JJ_1.1'
30      else if next_word = PositiveConjunctionType then
31        Insert *not* in front of curr_word // Method 'JJ_1.2'
32        Replace next_word with *nor* else
33        Insert *not* in front of curr_word // Method 'JJ_1.3'
34    else if curr_pos = IN then
35      Insert *not* in front of curr_word // Method 'IN_1.1'
36 Define method as method name if applicable edit occurs
37 return text, method, edit_id
```

TABLE I
EXAMPLE NEGATED CAUSAL SENTENCES PER METHOD

method	edit_text	alt_text	n
VB_1.2	Eyes with better vision at baseline had no more favorable prognosis, whereas eyes with initial macular detachment, intraoperative iatrogenic break, or heavy SO showed more unfavorable outcomes.	Eyes with better vision at baseline abstained a more favorable prognosis, whereas eyes with initial macular detachment, intraoperative iatrogenic break, or heavy SO showed more unfavorable outcomes.	35
VB_1.3	Age, female sex, BMI, non-HDL cholesterol, and polyps are not independent determinants for gallstone formation.	Age, female sex, BMI, non-HDL cholesterol, and polyps differ independent determinants for gallstone formation.	12
VB_1.4	Both general and central adiposity have no causal effects on CHD and type 2 diabetes mellitus.	Both general and central adiposity refuse causal effects on CHD and type 2 diabetes mellitus.	2
VB_2.1	Not "both a low-fat vegan diet and a diet based on ADA guidelines improved glycemic and lipid control in type 2 diabetic patients."	-	1
VB_3.1	Collectively, these findings did not indicate that energy-matched high intensity and moderate intensity exercise are effective at decreasing IHL and NAFLD risk that is not contingent upon reductions in abdominal adiposity or body mass.	Collectively, these findings contraindicate that energy-matched high intensity and moderate intensity exercise are effective at decreasing IHL and NAFLD risk that is not contingent upon reductions in abdominal adiposity or body mass.	174
VB_4.1	The benefits of exercise for reducing risk of chronic disease, including CVD, are well not known.	-	1
VB_5.1	A higher BMI and a greater prevalence of comorbidities had not driven patients to seek a more radical solution for their obesity, i.e., surgery.	A higher BMI and a greater prevalence of comorbidities had attract patients to seek a more radical solution for their obesity, i.e., surgery.	81
JJ_1.1	The effects of TRT on cardiovascular risk markers were not ambiguous.	-	6
JJ_1.2	Results are not encouraging nor demonstrate that exercise was popular and conveyed benefit to participants.	Results are discouraging and disprove that exercise was popular and conveyed benefit to participants.	15
JJ_1.3	While LSG weakens the LES immediately, it does not predictably not affect postoperative GERD symptoms; therefore, distensibility is not the only factor affecting development of postoperative GERD, confirming the multifactorial nature of post-LSG GERD.	While LSG weakens the LES immediately, it does not predictably impede postoperative GERD symptoms; therefore, distensibility is not the only factor affecting development of postoperative GERD, confirming the multifactorial nature of post-LSG GERD.	53
IN_1.1	Although further investigation of long-term and prospective studies is not needed, we identified four variables as predisposing factors for higher major amputation in diabetic patients through meta-analysis.	-	1

Example successful negations of causal sentences from CSci corpus. **method** refers to method label as per Algorithm 1. **edit_text** refers to direct negation from this Algorithm. **alt_text** refers to edit-alternate intervention using same negation location, but based off antonyms from WordNet, if available. Interventions, excluding lemmatisation or case-changes, are highlighted in green.

Algorithm 2: StrengthenRules – Causal strengthening scheme

Input: *edit_id, text_ids, text, pos, sentid2tid, curr_try=0***Output:** *text, method, edit_id*

```
1 Initialise ModalDict
2 curr_try  $\leftarrow$  curr_try + 1
3 curr_pos, curr_word  $\leftarrow$  pos[edit_id], text[edit_id]
4 next_pos, next_word  $\leftarrow$  pos[edit_id + 1], text[edit_id + 1] if valid else None
5 nnext_pos, nnext_word  $\leftarrow$  pos[edit_id + 2], text[edit_id + 2] if valid else None
6 while curr_try  $\leq$  max_try do
7   if lemma(next_word) = 'be' then
8     Replace curr_word with *was* // Method 'MOD_1.2'
9     Replace next_word with empty string
10  else if lemma(next_word) = 'have' then
11    if lemma(nnext_word) = 'be' then
12      Replace curr_word with *was* // Method 'MOD_3.2'
13      Replace next_word and nnext_word with empty string
14    else
15      Replace curr_word with *had* // Method 'MOD_3.1'
16      Replace next_word with empty string
17  else if curr_pos = MD next_pos = RB then
18    Replace curr_word with ModalDict[curr_word] // Method 'MOD_4.1'
19    Replace next_word with empty string
20  else
21    Replace curr_word with ModalDict[curr_word] // Method 'MOD_1.1'
22 Define method as method name if applicable edit occurs
23 return text, method, edit_id
```

TABLE II
EXAMPLE STRENGTHENED CONDITIONAL CAUSAL SENTENCES PER METHOD

method	edit_text	n
MOD_1_1	Physical therapy in conjunction with nutritional therapy may will help prevent weakness in HSCT recipients.	98
MOD_2_1	The rs7044343 polymorphism could be was involved in regulating the production of IL-33.	42
MOD_3_1	Increased titers of cows milk antibody before anti-TG2A and celiac disease indicates that subjects with celiac disease might have had increased intestinal permeability in early life.	21
MOD_4_1	Physical rehabilitation aimed at improving exercise tolerance can possibly will improve the long-term prognosis after operations for lung cancer.	13

Example successful strengthening of conditional causal sentences from CSci corpus. **method** refers to method label as per Algorithm 2, resulting in augments as per **edit_text**. Interventions, excluding lemmatisation or case-changes, are highlighted in green. Words removed from original version are striked out and highlighted in red.

TABLE III
DUPLICATE SENTENCES WITH DIFFERENT LABELS

S/N	Sentence	Label				Step
		c_0	c_1	c_2	c_3	
1	None the less, both artificially sweetened beverages and fruit juice were unlikely to be healthy alternatives to sugar sweetened beverages for the prevention of type 2 diabetes.	1		1		base
2	There was no effect on lumen volume, fibro-fatty and necrotic tissue volumes.	1	1			1to0
3	There are no indications that endogenous and exogenous gonadal hormones affect the radiation dose-response relationship.	1	1			1to0
4	In two randomized trials comparing the PCSK9 inhibitor bococizumab with placebo, bococizumab had no benefit with respect to major adverse cardiovascular events in the trial involving lower-risk patients but did have a significant benefit in the trial involving higher-risk patients.	1	1			1to0
5	Altering margin policies to follow either SSO-ASTRO or ABS guidelines would result in a modest reduction in the national re-excision rate.			1	1	2to1
6	Adding an allowance for accumulation of thyroidal iodine stores would produce an EAR of 72 ÅŽÅ¼g and a recommended dietary allowance of 80 ÅŽÅ¼g.			1	1	2to1
7	” In a randomized controlled trial of 230 infants with genetic risk factors for celiac disease, we did not find evidence that weaning to a diet of extensively hydrolyzed formula compared with cows milk-based formula would decrease the risk for celiac disease later in life.			1	1	2to1
8	However, there is no evidence that ABCB1 C3435T polymorphism would play a role in susceptibility to breast cancer in Morocco.	1	1			2to0

Sentences that had duplicates with differing labels. Rule-based de-duplication was performed, with the final label kept highlighted in green. Column ‘step’ refers to the step when edit merges with base, the duplicate appears. Do note that sentences S/N 7 and 8, to us, should be labelled c_0 , but was labelled as c_2 by original authors

TABLE IV
EXTENDED EXAMPLES OF COUNTERFACTUAL CAUSAL SENTENCE AUGMENTS

Variation	Sentence	Label
Original	TyG is effective to identify individuals at risk for NAFLD.	c_1
Regular (Edit)	TyG is not effective to identify individuals at risk for NAFLD.	c_0
Regular (edit-alternate)	TyG is ineffective to identify individuals at risk for NAFLD.	c_0
Shorten	TyG is ineffective	c_0
Multiples	is ineffective is ineffective is ineffective	c_0
Mask	[MASK] is ineffective to identify [MASK] at [MASK] for [MASK]	c_0
Synonyms	TyG exists inefficient to describe someone at take chances for NAFLD.	c_0
T5Paraphraser	Ineffective for identifying individuals at risk for NAFLD.	c_0

Example variations of two sentences from the CSci corpus fully augmented by rule-based algorithms. Interventions are highlighted in green.

TABLE V
NUMBER OF ITEMS PER CLASS LABEL

Edit_Conversion	Edit_Type	n_{c_0}	n_{c_1}	n_{c_2}	n_{c_3}	n
-	-	1356	494	213	998	3061
1to0	Regular	1356	491	212	995	3054
1to0	Shorten	1356	491	212	995	3054
1to0	Multiples	1356	491	212	995	3054
2to1	Regular	1353	494	209	995	3051
2to1	Shorten	1353	494	209	995	3051
2to1	Multiples	1353	494	209	995	3051
1to0, 2to1, 2to0	Regular	1356	494	209	995	3054
1to0, 2to1	Shorten, Regular	1356	494	209	995	3054

Number of items per class label after appending edits with base corpus, de-duplication and random sampling.

TABLE VI
BASE PREDICTED LABELS ON AUGMENTED SENTENCES

	1to0	2to1
pred	c_0	c_1
c_0	24	3
c_1	157	67
c_2	5	16
c_3	4	1
total	190	87

Count of predicted labels for not-causal (1to0) and stronger-causal (2to1) sentences when trained on original CSci corpus. Corresponds to results in Rows 1 and 5 of Table II respectively.

TABLE VII
PERFORMANCE METRICS OF BIOBERT MLP MODEL

Edit_Conversion	Edit_Type	P	R	F1	Acc	P _{Base}	R _{Base}	F1 _{Base}	Acc _{Base}
Yu et al. [12]		87.80	88.60	88.10	90.10	-	-	-	-
Ours (Base)		86.02	88.13	87.01	89.15	86.02	88.13	87.01	89.15
1to0	Regular	-1.81	-1.20	-1.55	-1.92	+0.29	-0.71	-0.19	-0.95
1to0	Shorten	+0.76	+1.45	+1.06	+0.89	+0.46	+0.78	+0.57	-0.04
1to0	Multiples	+1.47	+1.44	+1.46	+1.45	+1.05	+0.81	+0.93	+0.49
2to1	Regular	+1.96	+1.51	+1.75	+1.14	+0.98	+0.58	+0.80	+0.84
2to1	Shorten	+1.54	+0.54	+1.08	+0.91	+0.52	-0.29	+0.16	+0.62
2to1	Multiples	+1.51	+0.38	+0.98	+0.98	+0.53	-0.70	-0.05	+0.57
1to0, 2to1, 2to0	Regular	-1.22	-0.49	-0.91	-1.47	-0.18	-1.15	-0.64	-0.93
1to0, 2to1	Shorten, Regular	+2.98	+2.57	+2.80	+2.33	+1.90	+1.54	+1.73	+1.35

Performance of BioBERT MLP model trained on CSci corpus (Base plus edits), with edits matching existing labels and randomly sampled to match base class distribution. Results are for test set when trained and predicted over 5-folds. Precision (P), Recall (R), macro F-score (F1) and accuracy (Acc) are reported in percentages. Rows below “Ours (Base)” report relative changes to that row. Columns with lower script “Base” are calculated for base items only (i.e. performance for edits is ignored). The best performance per column is indicated by **boldface**.

TABLE VIII
PERFORMANCE METRICS OF BIOBERT MLP+SVM MODEL

Edit_Conversion	Edit_Type	P	R	F1	Acc	P _{Base}	R _{Base}	F1 _{Base}	Acc _{Base}
Ours (Base)		86.28	87.70	86.95	88.86	86.28	87.70	86.95	88.86
1to0	Regular	-2.72	-1.85	-2.33	-1.99	-0.89	-1.44	-1.18	-1.28
1to0	Shorten	+0.60	+1.36	+0.95	+1.19	+0.16	+0.67	+0.38	+0.18
1to0	Multiples	+1.18	+1.12	+1.14	+1.28	+0.68	+0.53	+0.60	+0.32
2to1	Regular	+0.97	+0.44	+0.73	+0.49	-0.14	-0.46	-0.28	+0.20
2to1	Shorten	+1.19	+0.54	+0.86	+1.08	+0.17	-0.65	-0.24	+0.71
2to1	Multiples	+0.92	+0.26	+0.62	+0.82	-0.21	-0.84	-0.50	+0.38
1to0, 2to1, 2to0	Regular	-1.26	-0.10	-0.73	-0.68	+0.09	-0.78	-0.39	-0.17
1to0, 2to1	Shorten, Regular	+1.25	+1.69	+1.45	+1.38	+0.00	+0.32	+0.14	+0.19

Performance of BioBERT MLP+SVM model trained on CSci corpus (Base plus edits), with edits matching existing labels and randomly sampled to match base class distribution. Results are for test set when trained and predicted over 5-folds. Precision (P), recall (R), macro F-score (F1) and accuracy (Acc) are reported in percentages. Rows below “Ours (Base)” report relative changes to that row. Columns with lower script “Base” are calculated for base items only (i.e. performance for edits is ignored). The best performance per column is indicated by **boldface**.