Causal Augmentation for Causal Sentence Classification

Fiona Anting Tan[1], Devamanyu Hazarika[2], See-Kiong Ng[1], Soujanya Poria[3] and Roger Zimmermann[2]

[1]Institute of Data Science, National University of Singapore

[2]School of Computing, National University of Singapore

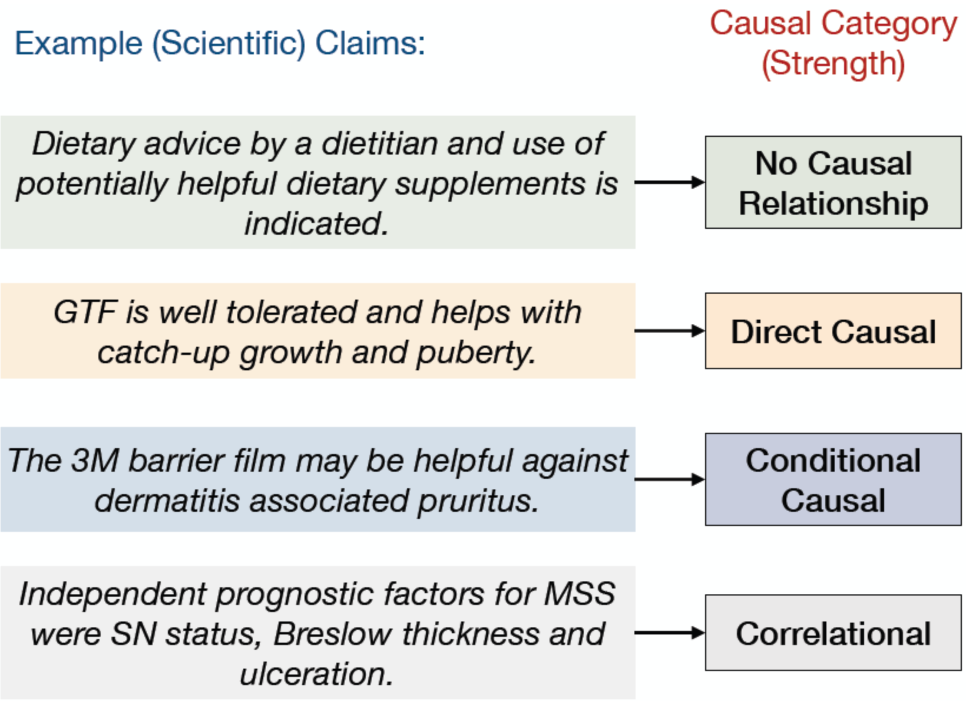
[3]Information Systems Technology and Design, Singapore University of Technology and Design

tan.f@u.nus.edu, hazarika@comp.nus.edu.sg, seekiong@nus.edu.sg, sporia@sutd.edu.sg, rogerz@comp.nus.edu.sg

TASK

01

Causal sentence classification (CSC) classifies textual claims into various categories of causal strengths.

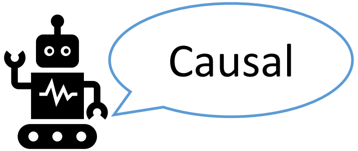


Dataset: PubMed-based **CSci** corpus (Yu et al., 2019)

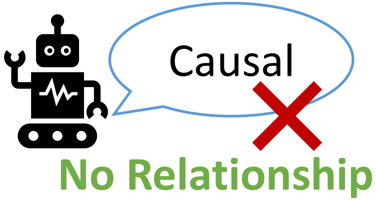
MOTIVATION

02 Current models are not robust to minimally perturbed sentences that differ in causal direction and strength.

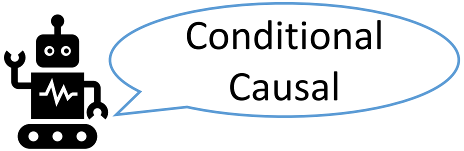
Both conbercept and ranibizumab are effective in the treatment of DME, achieving the similar clinical efficacy.



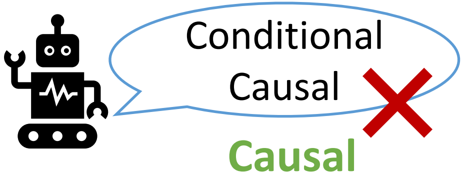
Both conbercept and ranibizumab are ~~effective~~ ineffective in the treatment of DME, achieving the similar clinical efficacy.



This suggests that TNP may play a role in enhancing wound healing.



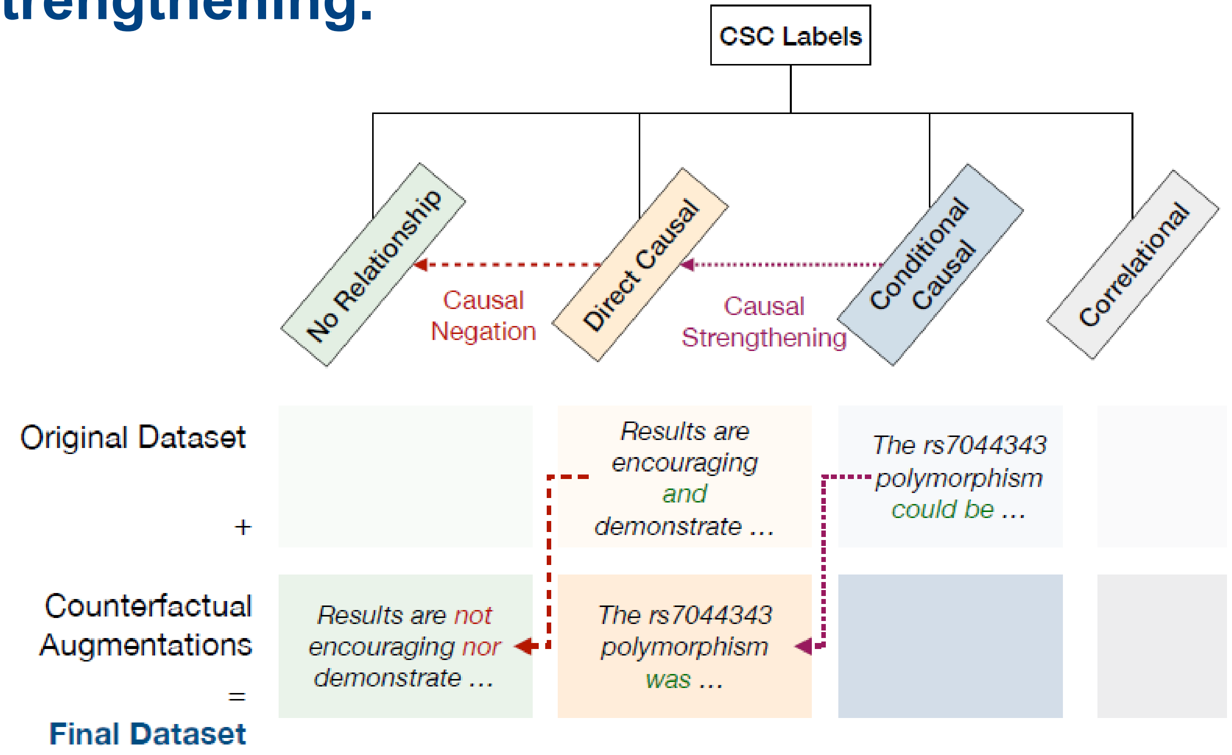
This suggests that TNP ~~may~~ will play a role in enhancing wound healing.



OUR APPROACH

03

Counterfactuals are generated purposefully for CSC by moving sentences across labels using Causal Negation and Strengthening.



04

In NEGATION, we negate the direction of causal statements from causal (c1) to no relationship (c0).

Method	REGULAR (EDIT)
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.
VB_1.3	Age, female sex, BMI, non-HDL cholesterol, and polyps are not independent determinants for gallstone formation.
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.
JJ_1.2	Results are not encouraging nor demonstrate that exercise was popular and conveyed benefit to participants.

(more templates shown in Appendix of paper)

05

In STRENGTHEN, we increased the strength of causal statements from conditional causal (c2) to causal (c1) by exploiting modal words.

Method	REGULAR (EDIT)
MOD_1.1	Physical therapy in conjunction with nutritional therapy ~~may~~ will help prevent weakness in HSCT recipients.
MOD_2.1	The rs7044343 polymorphism ~~could be~~ was involved in regulating the production of IL-33.
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.
MOD_4.1	Physical rehabilitation aimed at improving exercise tolerance ~~can possibly~~ will improve the long-term prognosis after operations for lung cancer.

06

We also experimented with other heuristics like shortening to a root phrase or multiplying key words.

Conversion	Edit Type	Sentence
NEGATION	Original	TyG is effective to identify individuals at risk for NAFLD.
	REGULAR (EDIT)	TyG is not effective to identify individuals at risk for NAFLD.
	REGULAR (EDIT-ALT)	TyG is ineffective to identify individuals at risk for NAFLD.
	SHORTEN	TyG is ineffective
	MULTIPLES	is ineffective is ineffective is ineffective
STRENGTHEN	Original	Moreover, TT genotype may reduce the risk of CAD in diabetic patients.
	REGULAR (Edit)	Moreover, TT genotype will reduce the risk of CAD in diabetic patients.

Table 1: Examples of counterfactual causal sentence augments. *Notes.* Interventions are highlighted in green. Causal Strengthening can also have SHORTEN and MULTIPLES edits but is excluded due to space constrains.

07 We experimented with two baseline models.

- BioBERT+MLP (MLP) (Yu et al., 2019)
- BioBERT+MLP+SVM (SVM)

$$
\begin{aligned}
z &= BERT(s), & z &\in \mathbb{R}^{h_1} && (1)\\
r &= MLP_1(z), & r &\in \mathbb{R}^{h_2} && (2)\\
o &= MLP_2(r), & o &\in \mathbb{R}^{c} && (3)\\
p &= SVM(r), & p &\in \mathbb{R}^{1}, && (4)
\end{aligned}
$$

$h_1 = 768$, $h_2 = 24$, and $c = 4$.

RESULTS

08 SOTA models are not robust to minimally altered sentences that change in causal direction or strength.

No Causal Relationship
Direct Causal
Conditional Causal
Correlational

Conversion	True Label	c_0	c_1	c_2	c_3	Total
NEGATION	c_0	24	157	5	4	190
STRENGTHEN	c_1	3	67	16	1	87

Table A6: Number of sentences predicted per class label for augmented dataset when trained on only original CSci corpus. *Notes*. Counts correspond to accuracy scores reported in Rows 1 and 3 of Table 3.

Conversion	n	MLP	SVM
Original	190	12.63	10.53
NEGATION	190	**+61.05**	**+62.63**
Original	87	77.01	73.56
STRENGTHEN	87	**+11.49**	**+13.79**

Table 3: Accuracy (in %) of BioBERT models trained on a subset of CSci corpus and predicted on a fully augmented difference set. *Notes*. The best performance per section per column is **bolded**.

09

Including a mixture of negated and strengthen edits improve model performance.

Conversion	Edit Type	MLP				SVM			
		F1	Acc	$F1_{Orig}$	Acc_{Orig}	F1	Acc	$F1_{Orig}$	Acc_{Orig}
Yu et al. (2019)		88.10	90.10	88.10	90.10	72.20	77.20	72.20	77.20
Ours (Base)		87.01	89.15	87.01	89.15	86.95	88.86	86.95	88.86
NEGATION	REGULAR	-1.55	-1.92	-0.19	-0.95	-2.33	-1.99	-1.18	-1.28
NEGATION	SHORTEN	+1.06	+0.89	+0.57	-0.04	+0.95	+1.19	+0.38	+0.18
NEGATION	MULTIPLES	+1.46	+1.45	+0.93	+0.49	+1.14	+1.28	+0.60	+0.32
STRENGTHEN	REGULAR	+1.75	+1.14	+0.80	+0.84	+0.73	+0.49	-0.28	+0.20
STRENGTHEN	SHORTEN	+1.08	+0.91	+0.16	+0.62	+0.86	+1.08	-0.24	**+0.71**
STRENGTHEN	MULTIPLES	+0.98	+0.98	-0.05	+0.57	+0.62	+0.82	-0.50	+0.38
NEGATION×SHORT, STRENGTHEN×REGU		**+2.80**	**+2.33**	**+1.73**	**+1.35**	+1.45	+1.38	+0.14	+0.19
NEGATION×MULTI, STRENGTHEN×REGU		+1.81	+1.35	+0.09	-0.10	**+1.95**	**+1.81**	**+0.62**	+0.61

10

Inclusion of edits during training can improve generalization in out-of-domain applications.

Conversion	Edit Type	SCITE (Li et al., 2021)				AltLex (Hidey and McKeown, 2016)			
		MLP		SVM		MLP		SVM	
		Acc	Acc_{Group}	Acc	Acc_{Group}	Acc	Acc_{Group}	Acc	Acc_{Group}
Ours (Base)		**86.28**	**85.83**	85.04	84.50	85.57	84.64	85.91	84.68
NEGATION	REGULAR	-1.46	-1.67	-0.36	-0.41	-0.22	-0.44	+0.18	+0.41
NEGATION	SHORTEN	-0.20	-0.27	+0.02	+0.02	+0.61	+0.54	+0.74	+1.05
NEGATION	MULTIPLES	-0.18	-0.16	-0.38	-0.38	+0.89	+0.95	**+1.19**	**+1.58**
STRENGTHEN	REGULAR	-0.27	-0.14	**+1.01**	**+1.10**	+0.51	+0.69	+0.54	+0.84
STRENGTHEN	SHORTEN	-3.40	-3.36	-0.11	-0.05	+0.30	+0.37	+0.99	+1.38
STRENGTHEN	MULTIPLES	-1.31	-1.28	-0.90	-0.90	+0.88	**+0.99**	+0.07	+0.29
NEGATION×SHORT, STRENGTHEN×REGU		-0.02	-0.05	+0.79	+0.63	**+0.94**	+0.84	+0.31	+0.41
NEGATION×MULTI, STRENGTHEN×REGU		-0.18	-0.16	+0.56	+0.56	+0.74	+0.88	+1.11	+1.33

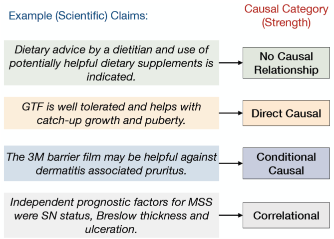
Thank you.

Fiona Anting Tan
tan.f@u.nus.edu
https://github.com/tanfiona/CausalAugment

Causal Augmentation for Causal Sentence Classification

Fiona Anting Tan[1], Devamanyu Hazarika[2], See-Kiong Ng[1], Soujanya Poria[3] and Roger Zimmermann[2]
[1]Institute of Data Science, National University of Singapore
[2]School of Computing, National University of Singapore
[3]Information Systems Technology and Design, Singapore University of Technology and Design
tan.f@u.nus.edu, hazarika@comp.nus.edu.sg, seekiong@nus.edu.sg, sporia@sutd.edu.sg, rogerz@comp.nus.edu.sg

Task



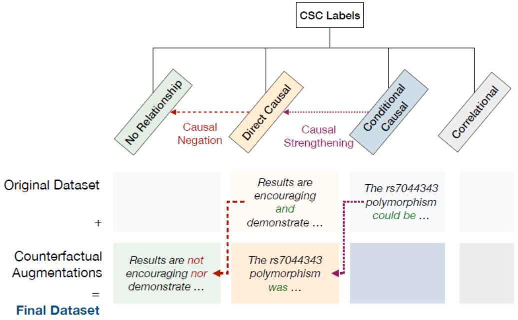
Causal sentence classification (CSC) classifies textual claims into various categories of causal strengths. Our main corpus is the PubMed-based CSci corpus (Yu et al., 2019) with four categories of causality.

Motivation

Conversion	n	MLP	SVM
Original	190	12.63	10.53
NEGATION	190	**+61.05**	**+62.63**
Original	87	77.01	73.56
STRENGTHEN	87	**+11.49**	**+13.79**

We found that models misclassify on augmented sentences that have been negated or strengthened with respect to its causal meaning. This is worrying since minor linguistic differences in causal sentences can have disparate meanings. For example, although the original MLP model only achieves 12.63% accuracy when predicting on negated examples, once we exposed the models to some negated examples during training, accuracy could increase to 73.68%.

Methodology



Conversion	Edit Type	Sentence
NEGATION	Original	TyG is effective to identify individuals at risk for NAFLD.
	REGULAR (EDIT)	TyG is not effective to identify individuals at risk for NAFLD.
	REGULAR (EDIT-ALT)	TyG is ineffective to identify individuals at risk for NAFLD.
	SHORTEN	TyG is ineffective
	MULTIPLES	is ineffective is ineffective is ineffective
STRENGTHEN	Original	Moreover, TT genotype may reduce the risk of CAD in diabetic patients.
	REGULAR (Edit)	Moreover, TT genotype will reduce the risk of CAD in diabetic patients.

Table 1: Examples of counterfactual causal sentence augments. *Notes.* Interventions are highlighted in green. Causal Strengthening can also have SHORTEN and MULTIPLES edits but is excluded due to space constrains.

We proposed generating augmented examples purposefully for CSC by moving sentences across labels using linguistic-based Causal Negation and Strengthening strategies. These augmentations were combined with the original dataset for model training. We also experimented with heuristics like shortening or multiplying root words of a sentence.

Results

I. Performance on CSci corpus

Conversion	Edit Type	MLP				SVM			
		F1	Acc	$F1_{Orig}$	Acc_{Orig}	F1	Acc	$F1_{Orig}$	Acc_{Orig}
Yu et al. (2019)		88.10	90.10	88.10	90.10	72.20	77.20	72.20	77.20
Ours (Base)		87.01	89.15	87.01	89.15	86.95	88.86	86.95	88.86
NEGATION	REGULAR	-1.55	-1.92	-0.19	-0.95	-2.33	-1.99	-1.18	-1.28
NEGATION	SHORTEN	+1.06	+0.89	+0.57	-0.04	+0.95	+1.19	+0.38	+0.18
NEGATION	MULTIPLES	+1.46	+1.45	+0.93	+0.49	+1.14	+1.28	+0.60	+0.32
STRENGTHEN	REGULAR	+1.75	+1.14	+0.80	+0.84	+0.73	+0.49	-0.28	+0.20
STRENGTHEN	SHORTEN	+1.08	+0.91	+0.16	+0.62	+0.86	+1.08	-0.24	**+0.71**
STRENGTHEN	MULTIPLES	+0.98	+0.98	-0.05	+0.57	+0.62	+0.82	-0.50	+0.38
NEGATION×SHORT, STRENGTHEN×REGU		**+2.80**	**+2.33**	**+1.73**	**+1.35**	+1.45	+1.38	+0.14	+0.19
NEGATION×MULTI, STRENGTHEN×REGU		+1.81	+1.35	+0.09	-0.10	**+1.95**	**+1.81**	**+0.62**	+0.61

II. Performance on Out-of-Domain datasets

Conversion	Edit Type	SCITE				AltLex			
		MLP		SVM		MLP		SVM	
		Acc	Acc_{Group}	Acc	Acc_{Group}	Acc	Acc_{Group}	Acc	Acc_{Group}
Ours (Base)		**86.28**	**85.83**	85.04	84.50	85.57	84.64	85.91	84.68
NEGATION	REGULAR	-1.46	-1.67	-0.36	-0.41	-0.22	-0.44	+0.18	+0.41
NEGATION	SHORTEN	-0.20	-0.27	+0.02	+0.02	+0.61	+0.54	+0.74	+1.05
NEGATION	MULTIPLES	-0.18	-0.16	-0.38	-0.38	+0.89	+0.95	**+1.19**	**+1.58**
STRENGTHEN	REGULAR	-0.27	-0.14	**+1.01**	**+1.10**	+0.51	+0.69	+0.54	+0.84
STRENGTHEN	SHORTEN	-3.40	-3.36	-0.11	-0.05	+0.30	+0.37	+0.99	+1.38
STRENGTHEN	MULTIPLES	-1.31	-1.28	-0.90	-0.90	+0.88	**+0.99**	+0.07	+0.29
NEGATION×SHORT, STRENGTHEN×REGU		-0.02	-0.05	+0.79	+0.63	**+0.94**	+0.84	+0.31	+0.41
NEGATION×MULTI, STRENGTHEN×REGU		-0.18	-0.16	+0.56	+0.56	+0.74	+0.88	+1.11	+1.33

Our strengthening schemes proved useful in improving model performance, while performance varies for negation regular edits. By including a mixture of edits when training, we achieved performance improvements beyond the baseline across both models, and within and out of corpus' domain, suggesting that our proposed augmentation can also help models generalize.