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

Abstract

Scarcity of annotated causal texts leads to poor robustness when training state-of-the-art language models for causal sentence classification. In particular, 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. Therefore, we propose the generation of counterfactual causal sentences by creating contrast sets (Gardner et al., 2020) to be included during model training. We experimented on two model architectures and predicted on two out-of-domain corpora. While our strengthening schemes proved useful in improving model performance, for negation, regular edits were insufficient. Thus, we also introduce heuristics like shortening or multiplying root words of a sentence. 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.

1 Introduction

Causality is an important concept for knowledge discovery as it conveys the idea of cause and effect. In the simplest sense, a causal relation exists between entities A and B through the statement "*A causes B*" or "*B is caused by A*". In recent years, causal relation extraction from text has garnered significant interest in Natural Language Processing (NLP) (Asghar, 2016; Xu et al., 2020; Yang et al., 2021).

Causal sentence classification (CSC) is the task of identifying sentences that contain causal meaning (Yu et al., 2019; Sumner et al., 2014; Mariko et al., 2020). Identification of causal sentences is often the first step in tasks like generating plot structures (Mirza and Tonelli, 2016a; Caselli and Vossen, 2017a) or constructing causal knowledge graphs (Heindorf et al., 2020) for further downstream Natural Language Understanding applications, like Question Answering (Dalal et al., 2021). Figure 1 demonstrates examples where similar claims are categorized by their causal strengths. CSC is challenging because the syntax of causality varies in context. Thus, it is difficult to exhaustively capture causal expressions, especially for implicit occurrences (Asghar, 2016). Negations and the absence of causality further complicate automatic causality identification (Heindorf et al., 2020).

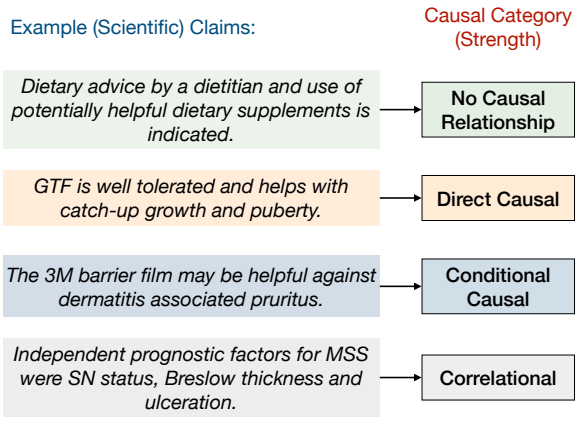

Figure 1: Causal sentence classification classifies textual claims into various categories of causal strengths.

Furthermore, there is a lack of good quality CSC datasets (Asghar, 2016; Xu et al., 2020). Most NLP datasets typically treat causal relation extraction as a subtask of relation extraction, where "Cause-Effect" is one of the many relation labels. However, we think that causality is a complex relation best learned using dedicated causal relation datasets. Such corpora that exist are mostly small in size (< 5000 sentences), except for AltLex (Hidey and McKeown, 2016) that has over 40000 sentences. Datasets also tend to label causal relations in an overly simplistic binary level (as 'causal' or 'not causal'). Only some works classify text by causal

strengths (Girju and Moldovan, 2002; Yu et al., 2019; Sumner et al., 2014).

Data augmentation is a natural avenue for handling small-sized datasets. Augments created must be meaningful to explain representation gaps in the current datasets. In causality, both the causal direction and strength matter. As such, we believe that models should be sensitive towards negations and semantics of words to avoid misclassification. For example, in Figure 1, the first three sentences include words related to *"help"*. However, the context of its usage and inclusion of modal words like *"may"* easily alters the intended causal strength of the sentence. This observation motivates us to artificially construct meaningful counterfactuals that would reflect the model's decision boundaries. We do so by applying rule-based schemes that negate causal relations or strengthen conditionally causal sentences. However, for negations, we notice that introducing edits is insufficient to improve model performance. Thus, we also explore adding heuristic edits.

We find that state-of-the-art (SOTA) language models, such as BERT (Devlin et al., 2019) with MLP or SVM classifiers, achieve improvements in classification performance when trained with our created counterfactuals. In addition, our evaluation on cross-domain datasets shows that training on augmented datasets (original plus edits) improves model generalization to out-of-domain (OOD) contexts. This is consistent with findings from (Kaushik et al., 2020a,b) in sentiment analysis and natural language inference contexts. In summary, we make the following contributions:

1. We show that current SOTA models are not robust to minimally perturbed sentences that differ in causal direction and strength. Therefore, we propose causal negation and strengthening schemes based on dependency and part-of-speech (POS) tags to augment causal sentences. To our knowledge, we are the first to study the effects of counterfactual augmentation in the context of causal claims classification.

2. We observe that simple heuristic edits on negated counterfactuals improve model effectiveness for the CSC task.

3. We show that a mixture of counterfactuals improves performance in the trained domain and also generalizes better to OOD corpora such as SCITE (Li et al., 2021) and AltLex (Hidey and McKeown, 2016).

Section 2 details related works in the literature and positions our work amongst them. Section 3 explains our methods for data augmentation, data processing and modeling. Section 4 presents and discusses our findings while Section 5 concludes.

2 Related Works

2.1 Causal Sentence Classification

Although causality is an important concept for knowledge discovery, benchmarking datasets and standardization of labeling rules have been limited, thus prohibiting empirical comparisons across methodologies (Asghar, 2016; Xu et al., 2020). Most NLP benchmarking datasets define causal relations as just one out of many class labels (e.g. Part-Whole) (Jurgens et al., 2012; Gábor et al., 2018; Caselli and Vossen, 2017b; Mirza et al., 2014; Mirza and Tonelli, 2016b). Others focus on causal relations and define such relations as a binary label (Li et al., 2021; Mariko et al., 2020; Hidey and McKeown, 2016). However, causality may not always occur at extremes in real-life statements, and correlation can get confused for causation (Buhse et al., 2018). As such, instead of using a binary model of causality, a better way is to classify varying "strengths" of causal relations in sentences. In fact, a seven-point scheme[1] was proposed by Sumner et al. (2014) to categorize causal statements from health-related news and academic press releases. Subsequently, Yu et al. (2019) adapted this for scientific texts into a four-level system. In this work, we adopt the four-level causality labeled corpus and classification model by Yu et al. (2019)[2].

There is also an often observed issue that NLP systems that perform well on task datasets do not generalize to "real-life scenarios", thereby misleading and overstating the accuracies and usefulness of their models. Ensuring model generalizability to other domains can be challenging. For example, Ramesh et al. (2012) showed discourse triggers are

[1]The seven levels of causal strengths are (1) no statement, (2) explicit statement of no relation, (3) correlational, (4) ambiguous (i.e., a relationship is present, but the direction and level is ambiguous), (5) conditional causal, (6) can cause, and (7) unconditionally causal.

[2]We were unable to work on Sumner et al.'s dataset as it was not publicly available and had very limited samples per class label.

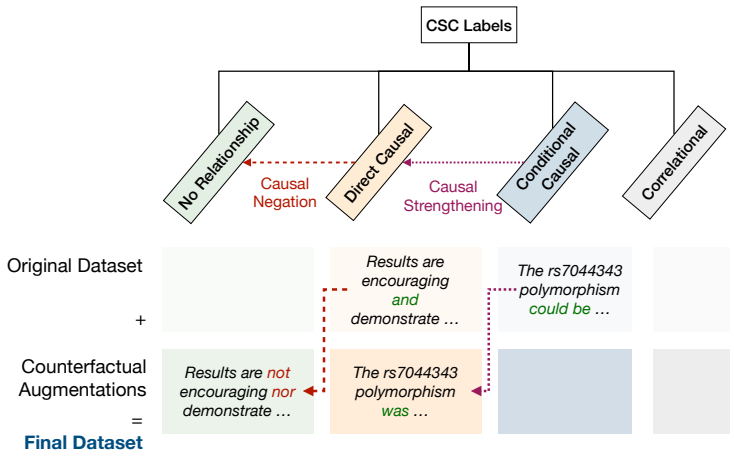

Figure 2: Strategies to generate counterfactual examples for CSC.

different between the biomedical and general domains. More focus has been placed on ensuring sufficient data representativeness and transferability of results onto OOD settings in recent years. In this work, we will also evaluate the generalizability of our models to classify causal sentences from other domains.

2.2 Counterfactuals in NLP

Counterfactual generation is a popular strategy for NLP researchers to test and improve model robustness via adversarial learning and attacks (Morris et al., 2020; Mahler et al., 2017) or for mitigating bias (Kaushik et al., 2020a; Maudslay et al., 2019).

Gardner et al. (2020) proposed using counterfactuals to fill local theoretical gaps in a model's decision boundary. They relied on expert judgments to generate similar but meaningfully different sentences and showed that SOTA models struggle on contrast sets compared to original test sets across multiple tasks. Recently, Wu et al. (2021) proposed a general-purpose counterfactual generator built on GPT-2 and also showed that the inclusion of realistic counterfactuals was useful across three different tasks. Their control codes included negation, delete, and restructure, amongst other options.[3] In our work, we generate counterfactuals purposefully for CSC, such as moving sentences across labels during Negation (*causal* $\rightarrow$ *no relationship*) and Strengthening (*conditional causal* $\rightarrow$ *causal*) strategies. We provide an automatic rule-based schema to negate and strengthen causal statements, focusing on precision over full coverage.[4]

Kaushik et al. (2020a) manually revised documents that would correspond to a counterfactual target label for sentiment analysis and natural language inference tasks. They showed that training with similar quantities of augmented data compared to the original improves generalization ability to OOD datasets. In this paper, we have also found that counterfactuals can help to improve model generalizability for CSC. Unlike their work, our linguistics-based augments do not rely on human intervention.

3 Methodology

3.1 Task Details

Our CSC task involved classifying a span of text with a causal label based on its intended meaning. We used the PubMed-based CSci corpus (Yu et al., 2019)[5] comprising of 3061 sentences, annotated with four levels of causal relation: *no relationship* (c_0), *causal* (c_1), *conditional causal* (c_2), and *correlational* (c_3).

3.2 Counterfactual Generation

In a low-resource setting, we propose creating counterfactuals that push causal sentences across labels to improve the robustness of models. Figure 2 demonstrates the two main strategies we employed to generate counterfactual examples for CSC: (1) Causal Negation ($c_1 \rightarrow c_0$) and (2) Causal

[3]Unfortunately, we did not investigate the negation and delete functions provided by Wu et al. (2021) but acknowledge this to be an important future work.

[4]Contemporaneously, we also contributed our rule-based algorithm to an open-source text augmentation effort at https://github.com/GEM-benchmark/NL-Augmenter under the transformation `negate_strengthen`.

[5]https://github.com/junwang4/causal-language-use-in-science

Strengthening ($c_2 \rightarrow c_1$). We discuss these strategies next.[6]

3.2.1 Causal Negation

In NEGATION, we negate the direction of causal statements from *causal* (c_1) to *no relationship* (c_0).

After obtaining POS tags and root words based on dependency trees[7], we performed negations around the root word. Our coding schema (Algorithm 1 in the Appendix) inserted negative words like *'no'*, *'not'*, *'nor'* or *'did not'* to negate the meaning of the sentence. 12 negation linguistic templates were used. Successfully negated sentences were termed as EDIT sentences. If no matching templates were found, the sentence was skipped. Of the 493 original (causal) sentences from the CSci corpus, 384 sentences had available negations.

To improve text flow, we used antonyms to replace negated edits where applicable. We did so by searching for antonyms of the original root word based on WordNet (Miller, 1995) and termed successful antonym edits as EDIT-ALT. To ensure a similar tense was used, we detected the original word's tense and applied the same tense onto the antonym word using the Pattern package (De Smedt and Daelemans, 2012). An example EDIT and EDIT-ALT sentence is shown in Table 1.

To decide between EDIT and EDIT-ALT versions, we calculated the Levenshtein edit distance of the original word versus the antonym. We selected EDIT-ALT only if the edit distance is less than or equal to 30% of the length of the longer word, rounded to the nearest integer. This allowed us to keep conversions like *'able'* $\rightarrow$ *'unable'* for more natural word flow, but discard bolder and more drastic changes like *'safe'* $\rightarrow$ *'dangerous'* and *'had'* $\rightarrow$ *'refused'* that suggested causality in the opposite direction (rather than *no relationship*) or were outright wrong. Finally, after dropping duplicates, we obtained 381 sentences that represented non-causality.

We were able to apply 11 out of the 12 linguistic templates to generate causal negation for the sentences in CSci. Most edits fell into the category where we negated the root verb or adjective of the sentence. Table A1 shows one randomly sampled example per available negation method when applied onto the CSci corpus. With respect to this table, Appendix A.1 briefly discusses the grammatical sanity of these sentences. We inspected these randomly sampled counterfactuals to verify that sentence flows were natural and desirable.

3.2.2 Causal Strengthening

For STRENGTHEN, we increased the strength of causal statements from *conditional causal* (c_2) to *causal* (c_1) by exploiting modal words. Similar to negation, we first obtained the POS tags and dependency trees for each sentence.

Algorithm 2 in the Appendix outlines the rule-based pseudo-code. To summarize, the 5 linguistic templates created converted modals based on the dictionary: *{'could', 'should', 'would'}* $\rightarrow$ *'would'* and *{'can', 'may', 'might', 'will'}* $\rightarrow$ *'will'*. If modals interacted with verbs with the lemma *'be'*, we replaced *'modal+be'* with *'was'* instead to convey certainty in the causal meaning. For special cases where the modal terms interacted with *'have'*, thereby forming conditional perfect tense, we converted the examples into simple past tense by replacing *'modal+have'* with *'had'*. When a modal was followed by an adverb (E.g. *"can possibly"*), the adverb was removed to avoid any deviation of the causal meaning from certainty.

Table A2 shows a randomly sampled example per causal strengthening method when applied onto the CSci corpus. Of the 213 available sentences, we successfully augmented 174 of them.

3.3 Dataset Processing

7 duplicated examples existed in the original CSci corpus and surfaced when we appended the edits with the original sentences. For such scenarios, we applied de-duplication based on priority rules discussed in Appendix A.2.

As our augmentations would increase the sample size for particular class labels, we randomly selected sentences to maintain the original class distribution. Our primary analysis focuses on randomly sampled datasets to eliminate the concern that any improved performance might result from increased data size or advantageous train set distribution.[8] As a side note, since the final dataset size is always slightly smaller than the original baseline

[6] Our edit schemes, model pipeline and augmented datasets are available at https://github.com/tanfiona/CausalAugment.

[7] We used NLTK (Wagner, 2010) to obtain POS tags in PennTreeBank format and spaCy (Honnibal et al., 2020) for dependency tree extraction.

[8] The aim of our paper is to demonstrate that any improvements in our scores are due to increased variations of examples per class label. These variations must be meaningful for any improvement in scores.

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.

due to the de-duplication step, the final distribution after random sampling slightly differs. The final sample counts across class labels per augmented dataset is reflected in Appendix Table A5.

3.4 Further Heuristics

Later in results Section 4.4.2, we observed that simple edits which highlight the main counterfactual phrase improved performance. Although these heuristics resulted in non-grammatical sentences, we believe that these edits explicitly emphasize augmented keywords for the model to learn the local syntactic changes better. Since we still trained the model with the original sentences (in fact, the majority), the model will not memorize on only non-grammatical examples.

An example sentence is detailed in Table 1 with the two heuristic options as follows:

- SHORTEN: We reduced the sentence length based on target/root word to cover a minimally interpretable phrase based on dependency parser. The final sentence might not be a consecutive slice from the original.

- MULTIPLES: We defined a phrase as one word before and after the target/root word (i.e. $PhraseLength = 3$). Phrases were then duplicated by a multiple of $OriginalSentenceLength/PhraseLength$ rounded to the nearest integer. This ensured that the final sentence was up to as long as the original length. Note that in the EDIT-ALT example of Table 1, *"is ineffective"* represents *"is not effective"*. Thus, although the actual phrase length was 2, the intended meaning is based off the latter phrase that had a length of 3. Hence, we maintained a fixed $PhraseLength$ for all sentences.

3.5 Out-of-domain Testing

In addition to training and validating on the CSci corpus, we also applied our trained models on two other datasets to demonstrate that exposing models to meaningful counterfactuals during training helps in OOD settings.

While the CSci corpus was constructed from scientific PubMed-based sentences, the SCITE (Li et al., 2021)[9] corpus comprised of general sentences extended from the SemEval 2010 Task 8 dataset (Hendrickx et al., 2010). On the other hand, AltLex (Hidey and McKeown, 2016)[10] contained sentences from English Wikipedia that included causal relations signaled by lexical markers. In AltLex, sentences can be duplicated if they have multiple relation markers and entities. Thus, we had to revise the corpus such that if a sentence had any causal relation, the sentence was labeled as *causal* and only one example was retained.

Additionally, because SCITE and AltLex have binary labels, we created two measures of accuracy. The first, 'Acc', considered only exact class labels (*no relationship* (c_0) and *causal* (c_1)) (i.e. predicting the other two labels is a misclassification). The second, 'Acc_{Group}', calculated accuracy after grouping [*no relationship, correlational*] into *no relationship* (c_0) and [*causal, conditional causal*] into *causal* (c_1) to align with the binary labels.

In total, we tested on 4439 sentences from SCITE and 37677 sentences from AltLex.

3.6 Modeling

In each setting, we trained and validated using $K = 5$ folds, with 5 epochs per fold. In both neural network set-ups, we used the standard cross-entropy loss for multi-class classification. For OOD testing, we took the majority prediction from

[9] https://github.com/Das-Boot/scite
[10] https://github.com/chridey/AltLex

the five trained models across the five folds. We implemented two models as follows:

3.6.1 BERT+MLP (MLP)

We replicated the best performing model on the CSci corpus (Yu et al., 2019) which was a BioBERT (Lee et al., 2020) plus multi-layer perceptron (MLP) pipeline. The default architecture used BioBERT embeddings fed through a single MLP layer serving as the classifier.

3.6.2 BERT+MLP+SVM (SVM)

Instead of applying LinearSVM based off unigrams and bigrams like the original authors (Yu et al., 2019), we believe a fairer comparison would be to use BERT embeddings as inputs into an SVM model. To allow for representation updates, for each sentence (s), the BioBERT encoder was first applied. Next, the BERT pooled output[11] (z) ran through two MLP layers (MLP_1 and MLP_2) to predict the class labels. After training, the second layer was discarded, and the hidden representation (r) was fed as fixed inputs into the SVM classifier. The equations below outlines this pipeline,

$$z = BERT(s), \qquad z \in \mathbb{R}^{h_1} \tag{1}$$

$$r = MLP_1(z), \qquad r \in \mathbb{R}^{h_2} \tag{2}$$

$$o = MLP_2(r), \qquad o \in \mathbb{R}^{c} \tag{3}$$

$$p = SVM(r), \qquad p \in \mathbb{R}^{1}, \tag{4}$$

where, p represents the final predicted label, and $h_1 = 768$, $h_2 = 24$, and $c = 4$.

4 Results & Discussion

4.1 Improvement over Baseline

Table 2 reports our performance on the CSci corpus. For the MLP baseline model, we were unable to exactly replicate the reported scores by Yu et al. (2019) of 90.1% accuracy and 88.1% macro F-score: We achieved slightly lower scores of 89.15% and 87.01% respectively. For SVM, our proposed implementation using updated BERT embeddings with a detached head was superior over Yu et al. (2019)'s unigram and bigrams method as we observed significant improvements of accuracy from 77.2% to 88.86% and macro F-score from 72.2% to 86.95%.

In our experiments, including a mixture of edits (NEGATION×SHORTEN with STRENGTHEN×REGULAR) during training returned the best performance across all metrics. Accuracy improved by 1.35% over our MLP baseline, achieving Acc_{Orig} of 90.60%.[12] Notice that we found improvements of accuracy and F-score beyond the original reported scores, even though our replicated scores were lower. The SVM model also demonstrated that including a mixture of edits during training improves performance, but in this setting, NEGATION×MULTIPLES with STRENGTHEN×REGULAR performed the best on average across metrics.

A possible explanation for our findings is that we successfully exposed our models to more sentence types of the real world. Furthermore, we intentionally created augments around label boundaries (i.e. the minor edits changes the sentences' labels). Therefore, the model learns better for the CSC task. Interestingly, for NEGATION, the heuristic edits improved performance against baseline more so than the REGULAR edits itself. Section 4.4.2 will expand on this finding.

4.2 Robustness on Edits

Table 3 highlights how current SOTA models are not robust to minimally altered sentences that changes in causal direction and strength.

To conduct the experiment, we randomly split the available negated edits (n=381) by half, keeping 191 negated sentences for training and the remaining 190 for testing. The 190 original sentences that correspond to the negated test set were removed from the original CSci corpus to avoid exposing models to highly similar sentences during training.[13] Models trained with this base train set dangerously predicted 157 out of 190 test sentences in the opposite direction as *causal* instead of *no relationship*. A shockingly dismal test accuracy of 12.63% was attained at best, and prediction counts are available in Appendix Table A6.

Our finding surfaces the problem that the models were likely memorizing key causal terms instead of understanding sentence structure and flow.

[11] Pooled output takes the hidden state from the first token.

[12] The full original set achieved 90.33% accuracy if we were to include the subset that is dropped out due to random sampling. To arrive at this value, we predicted the labels for this dropped-out subset like an OOD dataset, i.e. taken across 5-folds after training completes.

[13] In experiments not shown, the models trained on the full original CSci corpus almost certainly wrongly predicts the 190 negated sentences as *causal* To prove our point that models are memorizing causal terms, we removed the overlapping sentences to eliminate the possibility of the models memorizing similar sentences in train and test set instead.

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

Table 2: Performance on CSci corpus. *Notes.* BioBERT models trained on variations of CSci corpus (Original plus edits), with edits matching existing labels and randomly sampled to match base class distribution. Results are for validation set when trained and predicted over 5-folds. Macro F-score (F1) and accuracy (Acc) are in %. Columns with lowerscript "Orig" are calculated for original sentences only (i.e. Edits are ignored). Rows below "Ours (Base)" report relative changes to it. Best performance per column is **bolded**. Precision and Recall scores are available in Appendix Tables A7 and A8.

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**.

Therefore, they were unable to discern the negation involved. Inclusion of counterfactual examples helped to fill this representation gap. We created augmented sets by combining the base train set with the 191 negated train sentences for retraining. Once we exposed the models to these negated examples during training, the same models could predict the right label with up to 73.68% accuracy.

We also tested the models' efficacy on strengthened sentences converted from *conditional causal* to *causal*. Once counterfactual examples were included in the train set, improvements on test accuracy was obtained to a significant, but smaller, extent of +13.79% improvement at best.

4.3 Improving Generalization

In Table 4, we show that inclusion of edits during training also helps to improve generalization in cross-domain applications. Although our train dataset was an academic and scientific-based text represented by a BioBERT language model, we show that when we applied the same model to the general-based SCITE and Wikipedia-based AltLex corpora, inclusion of edits improved classification performance. For SCITE, we found improvements in generalization for the SVM model but not the MLP model. This could be due to our limited edit schemes that might not complement SCITE's sentence types. Nevertheless, for AltLex, consistent improvements for almost all edit combinations were obtained across both models. Overall, the mixture of edits with both conversion types once again reported the best average performance, demonstrating how such augments can indeed aid help models generalize.

4.4 Ablations

4.4.1 NEGATION vs. STRENGTHEN

While analyzing both result Tables 2 and 4, one might wonder why the REGULAR edit schemes helped improve performance for STRENGTHEN, but not for NEGATION conversions. One possible explanation for this phenomenon is as such – Since any sentence that did not represent any form of correlational or causal meaning falls under c_0, sentences that could fall under *no relationship* are lexically diverse. In other words, it is challenging to create edits that exhaustively reflect all c_0 sentence types. By and large, our negation schemes only covered one category of *no relationship* sentences, namely, sentences that imply *not causal*. On the other hand, *conditional causal* sentences were relatively well-defined in the original corpus. Therefore, STRENGTHEN did successfully represent most of the sentence types under c_2. Inclusion

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

Table 4: Performance on OOD datasets. *Notes.* BioBERT models trained on variations of CSci corpus (Original plus edits), with edits matching existing labels and randomly sampled to match base class distribution. For SCITE and AltLex, predictions are from takes mode class over 5-folds. Accuracies (Acc) are reported in %. Columns 'Acc' considers exact class labels, while 'Acc_{Group}' calculates accuracy after converting the four class labels into binary labels. Rows below "Ours (Base)" report relative changes to it. The best performance per column is **bolded**.

of these edits during training thus proved useful in highlighting the syntax that makes a sentence *causal* or *conditional causal* to the models.

4.4.2 Need for Heuristic Edits

Earlier in Table 2, we noted that models exposed to NEGATION×REGULAR edits were unable to effectively learn the label boundaries: Acc_{Orig} fell by 0.95% for the MLP model and 1.28% for the SVM compared to our baselines. However, when we performed simple heuristics like MULTIPLES, accuracy improved by +0.49% and +0.32% respectively. As for SHORTEN, accuracy rose by +0.18% for the SVM model, while the MLP model had a negligible reduction of -0.04%.

We study the net change in classification counts per model per label in Table 5 to explore this phenomenon. Given class labels i and j predicted by a model and our baseline respectively, we report the model's $NetChange_i = Right_i - Wrong_i = \sum_{j \neq i} n_{(i=true)j} - \sum_{i \neq j} n_{i(j=true)}$, where $i, j = c_0, c_1, c_2, c_3$ and n refers to the number of observations. $Right_i$ ($Wrong_i$) is the number of observations where a model predicts correctly (wrongly) for class label i but baseline predicts wrongly (correctly). When either MLP or SVM model is trained with the augmented NEGATION×REGULAR dataset, the model became confused and predicted poorly for *causal* (c_1) and *no relationship* (c_0) classes. Once the edits were presented in the heuristic forms, this situation improved.

We offer two plausible explanations for our findings: (1) It could be the case that highlighting the model to the short spans of (non-)causality aids its identification of the exact borders it needs to be sensitive to. (2) In the REGULAR form, non-causal sentences are linguistically very similar to causal ones. As mentioned in Section 4.4.1, these non-causal sentences only represent one out of many possible sentence types from c_0. Therefore, feeding some non-grammatical examples of c_0 might help make it more explicit to the model that c_0 can take a wide variety of sentences types. More work is needed to confirm either hypotheses.

Interestingly, we observed improvements in classification for labels we did not edit (c_3) in the majority of settings. This highlights the possibility that exposing models to minimally perturbed sentences around label boundaries might also improve comprehension beyond the introduced edits.

4.4.3 Capturing Causal Strengths

One benefit of capitalizing on CSci's four-label format is that our methodology is now able to identify causal strengths in SCITE and AltLex corpora beyond the original binary labels. For SCITE, the baseline MLP model originally labeled five sentences as *conditional causal*. When training the model with STRENGTHEN×REGULAR edits, four remained as *conditional causal* (c_2) while one of the sentence[14] correctly switched label to *causal* (c_1). For the baseline SVM model, seven sentences were tagged as c_2, of which four remained, and the same one as MLP's converted to c_1. One[15] cor-

[14] "In the present recession, which has been triggered by a collapse in land prices, land-value taxation would reverse the collapse - not by re-inflating a temporary speculative bubble, but by inducing investment in infrastructure that permanently enhances the utility of the land."

[15] "The glass tealight holder appears to float inside the metal spiral as it spins in the gentle breeze."

Conversion	Edit Type	MLP					SVM				
		c_0	c_1	c_2	c_3	**Total**	c_0	c_1	c_2	c_3	**Total**
NEGATION	REGULAR	-13	-18	+10	0	-21	-9	-21	+1	-2	-31
NEGATION	SHORTEN	-15	+9	+9	+2	+5	-4	+7	+1	+6	+10
NEGATION	MULTIPLES	-5	+9	+5	+9	+18	-1	+8	+3	+3	+13
STRENGTHEN	REGULAR	-7	+12	+10	+9	+24	+1	+1	+5	-4	+3
STRENGTHEN	SHORTEN	-7	+11	+6	+6	+16	0	+8	+3	+5	+16
STRENGTHEN	MULTIPLES	-14	+13	+7	+10	+16	-12	+11	+6	+3	+8
NEGATION×SHORT, STRENGTHEN×REGU		+2	+20	+10	+9	+41	-2	+12	+1	-4	+7
NEGATION×MULTI, STRENGTHEN×REGU		-16	+6	+5	+7	+2	+2	+9	+2	+5	+18

Table 5: Net change in correct classification counts on CSci corpus compared to "Ours (Base)" for original examples. Note that NEGATION is the conversion of $c_1 \rightarrow c_0$ and STRENGTHEN is the conversion of $c_2 \rightarrow c_1$;

rectly switched to *no relationship* (c_0) as labeled, while the last sentence[16] converted to *correlational* (c_3), which is surprising because we did not edit any sentences to or from class c_3. Unfortunately, the authors of SCITE tagged this sentence as causal, which means this is considered to be mislabeled. However, the sentence contains signals like *'corresponds to'*, which we believe should be correlational, not causal. Our short qualitative analysis again supports the earlier quantitative study that exposing models to meaningfully augmented sentences across labels could improve classification even for the other uninvolved labels.

4.4.4 Other Experiments

We also explored other popular methodologies but did not obtain consistent and significant improvements from baseline. These include, *(i)* creating more edit types (using masking, synonyms and paraphrasers), *(ii)* extending to a five-way classification problem (by labelling negated edits as a new class label representing *not causal*, separate from *no relationship* (c_0)), and *(iii)* experimenting with some contrastive learning loss functions. Appendix Section A.3 details these experiments further for interested readers.

5 Conclusion & Future Work

We explored the task of CSC in a low-resource setting. Following recent literature, we generated counterfactual sentences via rule-based edits that change sentences' causal direction and strength. We showed that SOTA CSC models worryingly misclassifies on such augmented sentences. This concern can be mitigated by including of our edits during training. We demonstrated that our proposal improves classification performance both on original and edit sentences, and within and outside of the corpus' domain. However, for NEGATION, we found that the regular format was insufficient to teach effective decision boundaries given limited data size and augmentation templates. Therefore, proposed heuristic edits and found performance improvements for both training and OOD contexts.

For future work, Yu et al. (2020)'s recent corpus using scientific press statements annotated with the same four class labels of causality is a promising dataset to replicate our findings upon. Additionally, we utilized rule-based augment schemes which have a finite number of working templates. Thus, our augmentations might not be lexically diverse. Therefore, our subsequent steps would be to explore SOTA NLP augmentation and generation tools, like from Wu et al. (2021) and Ross et al. (2021). Furthermore, it might be worthwhile to find alternative models that can learn directly from the augmented datasets without the need for heuristics.

Lastly, our work did not go beyond the "correctness" of the claims. However, in reality, one has to distinguish between causal effects as factual events of real-world or at the level of "meta-causality" (Andersson et al., 2020). Hence, grounding the claims to world knowledge will be an important research avenue to pursue.

Acknowledgements

This research is supported by Singapore Ministry of Education Academic Research Fund Tier 1 under MOE's official grant number T1 251RES2029.

[16] "The increase of the signal might correspond to formation of the high-density excitons, while the reduction of the signal originates from the relaxation."

References

Marta Andersson, Murathan Kurfalı, and Robert Östling. 2020. A sentiment-annotated dataset of English causal connectives. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 24–33, Barcelona, Spain. Association for Computational Linguistics.

Nabiha Asghar. 2016. Automatic extraction of causal relations from natural language texts: A comprehensive survey. *CoRR*, abs/1605.07895.

Susanne Buhse, Anne Christin Rahn, Merle Bock, and Ingrid Mühlhauser. 2018. Causal interpretation of correlational studies–analysis of medical news on the website of the official journal for german physicians. *PLoS One*, 13(5):e0196833.

Tommaso Caselli and Piek Vossen. 2017a. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017b. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 77–86. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Dhairya Dalal, Mihael Arcan, and Paul Buitelaar. 2021. Enhancing multiple-choice question answering with causal knowledge. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 70–80, Online. Association for Computational Linguistics.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *J. Mach. Learn. Res.*, 13(1):2063–2067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 679–688. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 1307–1323. Association for Computational Linguistics.

Roxana Girju and Dan I. Moldovan. 2002. Text mining for causal relations. In *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, May 14-16, 2002, Pensacola Beach, Florida, USA*, pages 360–364. AAAI Press.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 3023–3030. ACM.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

David Jurgens, Saif Mohammad, Peter D. Turney, and Keith J. Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Mon-*

tréal, Canada, June 7-8, 2012, pages 356–364. The Association for Computer Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020a. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Divyansh Kaushik, Amrith Setlur, Eduard H. Hovy, and Zachary C. Lipton. 2020b. Explaining the efficacy of counterfactually-augmented data. *CoRR*, abs/2010.02114.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, 423:207–219.

Taylor Mahler, Willy Cheung, Micha Elsner, David King, Marie-Catherine de Marneffe, Cory Shain, Symon Stevens-Guille, and Michael White. 2017. Breaking NLP: Using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 33–39, Copenhagen, Denmark. Association for Computational Linguistics.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (FinCausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32, Barcelona, Spain (Online). COLING.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5266–5274. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2016a. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75, Osaka, Japan. The COLING 2016 Organizing Committee.

Paramita Mirza and Sara Tonelli. 2016b. CATENA: causal and temporal relation extraction from natural language texts. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 64–75. ACL.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 119–126. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Balaji Polepalli Ramesh, Rashmi Prasad, Tim Miller, Brian Harrington, and Hong Yu. 2012. Automatic discourse connective detection in biomedical text. *J. Am. Medical Informatics Assoc.*, 19(5):800–808.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E. Peters, and Matt Gardner. 2021. Tailor: Generating and perturbing text with semantic controls. *CoRR*, abs/2107.07150.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D Chambers. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*, 349.

Wiebke Wagner. 2010. Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit - o'reilly media, beijing, 2009, ISBN 978-0-596-51649-9. *Lang. Resour. Evaluation*, 44(4):421–424.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Jinghang Xu, Wanli Zuo, Shining Liang, and Xianglin Zuo. 2020. A review of dataset and labeling methods for causality extraction. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1519–1531. International Committee on Computational Linguistics.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2021. A survey on extraction of causal relations from natural language text. *CoRR*, abs/2101.06426.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4663–4673. Association for Computational Linguistics.

Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. Measuring correlation-to-causation exaggeration in press releases. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4860–4872. International Committee on Computational Linguistics.

A Appendix

A.1 Negation Examples

Appendix Table A1 shows one randomly sampled example per available negation method when applied onto the CSci corpus. As shown, most examples fell into 'VB_3.1', 'VB_5.1', 'JJ_1.3' and 'VB_1.2' types, for which the templates in Algorithm 1 worked well for[17]. For rarer method types, like 'VB_2.1', the templates seemed 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 was unnatural but not grammatically wrong.

A.2 De-duplication

After appending original sentences with edits, we conducted de-duplication. Appendix Table A3 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 one (sentence S/N 1) was duplicated with different labels (c_0 and c_2). We manually changed this to retain only the c_0-labeled example. The total data size thus reduced from 3061 to 3054. This explains the differences in data size and distribution when comparing the original versus augmented sets shown in Table A5. We also take this chance to highlight concerns that some sentences in CSci were labeled 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 was that if an edit was performed but remained identical to the original, the original must have been mislabeled. 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.

[17] We 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, e.g. *'should'*).

A.3 Other Experiments

Other experiments that were conducted but did not produce significant improvements are mentioned here.

Other Edit Types Three were explored:

- MASK: Based on POS, all nouns were replaced by the token *"[MASK]"*.
- SYNONYMS: Using WordNet synonyms, we skipped common words[18] and randomly substituted up to 5 words. Synonyms matched the tense and plurarity of original words using Pattern package, which we note, had imperfections.
- T5PARA: We ran the sentence through a pretrained T5-paraphraser model[19] to generate paraphrased sentences.

Appendix Table A4 shows an example sentence with the above edits for the same causal sentence of Table 1. With the SVM model, only STRENGTHEN×SYNONYMS appended with original increased accuracy on CSci by 1.01% while STRENGTHEN×T5PARA increased accuracy by 0.39%. However, these findings could not be replicated across to the MLP model nor for NEGATION.

Extending to a Five-way Classification In our main set up, we focused on edits that matched the original labels and were randomly sampled such that the unified train set matches base class distribution for fairer comparison to baseline. Successful NEGATION examples were labeled *no relationship* (c_0). However, to the extent that we believe negated causal statements deserve a class of their own, we also explore the event when negations were labeled with a new level *not causal* (c_4) instead. Based on the set up for Table 3, we obtained even higher improvements in accuracy of +70.53% and +74.74% for the MLP and SVM model respectively. This could be due to the clearer distinction of a *not causal* sentence structure compared to if we were to combine them with other *no relationship* statements. When we extended the MLP and SVM model to work with such a five-way classification set up, we did observe improvements in Acc_{Orig}

[18] We do not try to find synonyms for common words with these POS types: 'DT','IN', 'EX', 'CC', 'MD', 'WP', 'WD', 'WR', 'UH', 'RP', 'SY', 'PO'

[19] https://huggingface.co/ramsrigouthamg/t5_paraphraser

for SHORTEN, MULTIPLES and SYNONYMS edit types. However, because we cannot truly balance the dataset (random sampling does not apply here because we have a whole new class), we cannot be certain if the improvements were due to the larger dataset or the model picking up on the boundaries. Furthermore, the improvements did not generalize on our OOD set ups.

Other Training Setups In addition to standard cross-entropy based supervised learning, we also explored contrastive learning schemes. In particular, we trained with Supervised Contrastive Loss (SupCon) (Khosla et al., 2020; Chen et al., 2020) and Triplet Margin Loss (Paszke et al., 2019). In the contrastive setup, we introduced counterfactuals as the negative examples for each anchor sentence. For positive samples, we used SHORTEN, SYNONYMS and T5PARA augmentation strategies derived from the original anchor sentence. However, our results did not provide performance improvements in either CSci or OOD datasets, highlighting the challenge in building a generalized scheme of counterfactual generations. Exploring avenues in contrastive learning remains a critical future work.

A.4 Reproducibility Checklist

We include additional details about our main experiment not highlighted in other parts of the paper.

- **Computing Infrastructure:** Tesla V100 SXM2 32 GB
- **MLP Hyperparameters:** “attention_probs_dropout_prob": 0.1, “hidden_act": “gelu", “hidden_dropout_prob": 0.1, “hidden_size": 768, “initializer_range": 0.02, “intermediate_size": 3072, “layer_norm_eps": 1e-12, “max_position_embeddings": 512, “num_attention_heads": 12, “num_hidden_layers": 12, “type_vocab_size": 2, “vocab_size": 28996
- **SVM Hyperparameters:** kernel: “linear", “C": 1e-2
- **Average Runtime:** For 5 epochs and 5 folds, our baseline MLP model took approximately 22 minutes 51 seconds to train and validate for the CSci dataset.

A.5 Additional Figures & Tables

Algorithm 1: NegationRules – Causal negation scheme

Input: $edit\_id$, $text\_ids$, $text$, pos, $sentid2tid$, max_try=2, $curr\_try$=0
Output: $text, method, edit\_id$

$curr\_try \leftarrow curr\_try + 1$
$curr\_pos, curr\_word \leftarrow pos[edit\_id], text[edit\_id]$
$prev\_pos, prev\_word \leftarrow pos[edit\_id - 1], text[edit\_id - 1]$ if valid else None
$next\_pos, next\_word \leftarrow pos[edit\_id + 1], text[edit\_id + 1]$ if valid else None
while $curr\_try <= max\_try$ **do**
  if $curr\_pos = VB$ **then**
    if $curr\_word = AuxilliaryType$ **then**
      if $edit\_id = max(text\_ids)$ **then**
        Insert *not* in front of $curr\_word$ // Method "VB_1.1'
      else if $next\_word = DeterminerType$ **then**
        Replace $next\_word$ with *no* // Method 'VB_1.2'
        $edit\_id \leftarrow edit\_id + 1$
      else if $next\_word = NounType$ **then**
        Insert *not* behind of $curr\_word$ // Method 'VB_1.3'
      else if $next\_pos = VB$ **then**
        Insert *no* behind of $curr\_word$ // Method 'VB_1.4'
    else if $edit\_id = min(text\_ids)$ **then**
      Replace $curr\_word$ with *Not* + lowercased $curr\_word$ // Method 'VB_2.1'
    else if $prev\_word = NounType$ **then**
      Replace $curr\_word$ with *did not* + $lemma(curr\_word)$ // Method 'VB_3.1'
    else if $edit\_id = max(text\_ids)$ **then**
      Insert *not* in front of $curr\_word$ // Method 'VB_4.1'
    else if $prev\_word = AuxilliaryType\ next\_pos = IN|TO$ **then**
      Insert *not* in front of $curr\_word$ // Method 'VB_5.1'
  else if $curr\_pos = NN$ **then**
    Get $head\_id$ of head word of $curr\_word$ based on dependency tree
    $text, method, edit\_id \leftarrow$ NegationRules($head\_id$, $text\_ids$, $text$, pos, $sentid2tid$, $curr\_try$)
  else if $curr\_pos = JJ$ **then**
    if $edit\_id = max(text\_ids)$ **then**
      Insert *not* in front of $curr\_word$ // Method 'JJ_1.1'
    else if $next\_word = PositiveConjuctionType$ **then**
      Insert *not* in front of $curr\_word$ // Method 'JJ_1.2'
    Replace $next\_word$ with *nor* **else**
      Insert *not* in front of $curr\_word$ // Method 'JJ_1.3'
  else if $curr\_pos = IN$ **then**
    Insert *not* in front of $curr\_word$ // Method 'IN_1.1'
Define $method$ as method name if applicable edit occurs
return $text, method, edit\_id$

Method	REGULAR (EDIT)	REGULAR (EDIT-ALT)	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

Table A1: Example negated causal sentences per method *Notes.* "Method" refers to NEGATION method label as per Algorithm 1. REGULAR (EDIT) refers to direct negation from this Algorithm. REGULAR (EDIT-ALT) refers to alternate intervention using same negation location, but based off antonyms from WordNet, if available. Interventions, excluding lemmatization or case-changes, are highlighted in green. "n" is the number of successful conversions applicable in CSci corpus.

Algorithm 2: StrengthenRules – Causal strengthening scheme

Input: $edit\_id$, $text\_ids$, $text$, pos, $sentid2tid$, $curr\_try$=0
Output: $text, method, edit\_id$

Initialize $ModalDict$
$curr\_try \leftarrow curr\_try + 1$
$curr\_pos, curr\_word \leftarrow pos[edit\_id], text[edit\_id]$
$next\_pos, next\_word \leftarrow pos[edit\_id + 1], text[edit\_id + 1]$ if valid else None
$nnext\_pos, nnext\_word \leftarrow pos[edit\_id + 2], text[edit\_id + 2]$ if valid else None
while $curr\_try <= max\_try$ **do**
    if $lemma(next\_word) =$ 'be' **then**
        Replace $curr\_word$ with *was* // Method 'MOD_1.2'
        Replace $next\_word$ with empty string
    else if $lemma(next\_word) =$ '$have$' **then**
        if $lemma(nnext\_word) =$ 'be' **then**
            Replace $curr\_word$ with *was* // Method 'MOD_3.2'
            Replace $next\_word$ and $nnext\_word$ with empty string
        else
            Replace $curr\_word$ with *had* // Method 'MOD_3.1'
            Replace $next\_word$ with empty string
    else if $curr\_pos = MD$ & $next\_pos = RB$ **then**
        Replace $curr\_word$ with $ModalDict[curr\_word]$ // Method 'MOD_4.1'
        Replace $next\_word$ with empty string
    else
        Replace $curr\_word$ with $ModalDict[curr\_word]$ // Method 'MOD_1.1'
Define $method$ as method name if applicable edit occurs
return $text, method, edit\_id$

Method	REGULAR (EDIT)	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

Table A2: Example strengthened conditional causal sentences per method. *Notes.* "Method" refers to strengthening method label as per Algorithm 2, resulting in augments as per REGULAR (EDIT). Interventions, excluding lemmatization or case-changes, are highlighted in green. Words removed from original version are striked out and highlighted in red. "n" is the number of successful conversions applicable in CSci corpus.

S/N	Sentence	Label				Conversion
		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		Original
2	There was no effect on lumen volume, fibro-fatty and necrotic tissue volumes.	1	1			NEGATION
3	There are no indications that endogenous and exogenous gonadal hormones affect the radiation dose-response relationship.	1	1			NEGATION
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			NEGATION
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		STRENGTHEN
6	Adding an allowance for accumulation of thyroidal iodine stores would produce an EAR of 72 ÃŽÂ¼g and a recommended dietary allowance of 80 ÃŽÂ¼g.		1	1		STRENGTHEN
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		STRENGTHEN

Table A3: Sentences that had duplicates with differing labels. *Notes.* Rule-based de-duplication was performed, with the final label kept highlighted in green. "Conversion" refers to the augmented edit dataset that when we merge with the original, the duplicate appears. Do note that Sentence S/N 7, to us, should be labeled as *no relationship* (c_0), but was labeled as *conditional causal* (c_2) by original authors.

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
	MASK	[MASK] is ineffective to identify [MASK] at [MASK] for [MASK]
	SYNONYMS	TyG exists inefficient to describe someone at take chances for NAFLD.
	T5PARA	Ineffective for identifying individuals at risk for NAFLD.

Table A4: Extended examples of counterfactual causal sentence augments *Notes.* Interventions are highlighted in green.

Conversion	Edit Type	n_c_0	n_c_1	n_c_2	n_c_3	n
Original (Yu et al., 2019)		1356	494	213	998	3061
NEGATION	REGULAR	1356	491	212	995	3054
NEGATION	SHORTEN	1356	491	212	995	3054
NEGATION	MULTIPLES	1356	491	212	995	3054
STRENGTHEN	REGULAR	1353	494	209	995	3051
STRENGTHEN	SHORTEN	1353	494	209	995	3051
STRENGTHEN	MULTIPLES	1353	494	209	995	3051
NEGATION×SHORT, STRENGTHEN×REGU		1356	494	209	995	3054
NEGATION×MULTI, STRENGTHEN×REGU		1356	494	210	995	3055

Table A5: Number of sentences per class label after appending edits with base corpus, de-duplication and random sampling. Note that the dataset corresponding to the first row did not undergo de-duplication (i.e. we used the original corpus as is).

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	Edit Type	P	R	F1	Acc	P_{Orig}	R_{Orig}	$F1_{Orig}$	Acc_{Orig}
Yu et al. (2019)		87.80	88.60	88.10	90.10	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
NEGATION	REGULAR	-1.81	-1.20	-1.55	-1.92	+0.29	-0.71	-0.19	-0.95
NEGATION	SHORTEN	+0.76	+1.45	+1.06	+0.89	+0.46	+0.78	+0.57	-0.04
NEGATION	MULTIPLES	+1.47	+1.44	+1.46	+1.45	+1.05	+0.81	+0.93	+0.49
STRENGTHEN	REGULAR	+1.96	+1.51	+1.75	+1.14	+0.98	+0.58	+0.80	+0.84
STRENGTHEN	SHORTEN	+1.54	+0.54	+1.08	+0.91	+0.52	-0.29	+0.16	+0.62
STRENGTHEN	MULTIPLES	+1.51	+0.38	+0.98	+0.98	+0.53	-0.70	-0.05	+0.57
NEGATION×SHORT, STRENGTHEN×REGU		**+2.98**	**+2.57**	**+2.80**	**+2.33**	**+1.90**	**+1.54**	**+1.73**	**+1.35**
NEGATION×MULTI, STRENGTHEN×REGU		+1.72	+1.91	+1.81	+1.35	-0.02	+0.23	+0.09	-0.10

Table A7: Performance of BERT+MLP on CSci corpus. *Notes.* BioBERT models trained on variations of CSci corpus (Original 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 %. Columns with lowerscript "Orig" are calculated for base items only (i.e. Edits are ignored). Rows below "Ours (Base)" report relative changes to it. The best performance per column is **bolded**.

Conversion	Edit Type	P	R	F1	Acc	P_{Orig}	R_{Orig}	$F1_{Orig}$	Acc_{Orig}
Yu et al. (2019)		73.90	71.10	72.20	77.20	73.90	71.10	72.20	77.20
Ours (Base)		86.28	87.70	86.95	88.86	86.28	87.70	86.95	88.86
NEGATION	REGULAR	-2.72	-1.85	-2.33	-1.99	-0.89	-1.44	-1.18	-1.28
NEGATION	SHORTEN	+0.60	+1.36	+0.95	+1.19	+0.16	**+0.67**	+0.38	+0.18
NEGATION	MULTIPLES	+1.18	+1.12	+1.14	+1.28	+0.68	+0.53	+0.60	+0.32
STRENGTHEN	REGULAR	+0.97	+0.44	+0.73	+0.49	-0.14	-0.46	-0.28	+0.20
STRENGTHEN	SHORTEN	+1.19	+0.54	+0.86	+1.08	+0.17	-0.65	-0.24	**+0.71**
STRENGTHEN	MULTIPLES	+0.92	+0.26	+0.62	+0.82	-0.21	-0.84	-0.50	+0.38
NEGATION×SHORT, STRENGTHEN×REGU		+1.25	**+1.69**	+1.45	+1.38	+0.00	+0.32	+0.14	+0.19
NEGATION×MULTI, STRENGTHEN×REGU		**+2.23**	+1.62	**+1.95**	**+1.81**	**+0.89**	+0.29	**+0.62**	+0.61

Table A8: Performance of BERT+MLP+SVM on CSci corpus. *Notes.* Yu et al.'s SVM method does not use BERT inputs. Our BioBERT models are trained on variations of CSci corpus (Original 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 %. Columns with lowerscript "Orig" are calculated for base items only (i.e. Edits are ignored). Rows below "Ours (Base)" report relative changes to it. The best performance per column is **bolded**.