The Causal News Corpus: Annotating Causal Relations in Event Sentences from News

Fiona Anting Tan[1], Ali Hürriyetoğlu[2], Tommaso Caselli[3], Nelleke Oostdijk[4],
Tadashi Nomoto[5], Hansi Hettiarachchi[6], Iqra Ameer[7], Onur Uca[8],
Farhana Ferdousi Liza[9], Tiancheng Hu[10]

[1] Institute of Data Science, National University of Singapore, Singapore, [2] Koc University, Turkey,
[3] Rijksuniversiteit Groningen, Netherlands, [4] Radboud University, Netherlands,
[5] National Institute of Japanese Literature, Japan, [6] Birmingham City University, United Kingdom,
[7] Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico,
[8] Department of Sociology, Mersin University, Turkey, [9] University of East Anglia, United Kingdom,
[10] ETH Zürich, Switzerland
tan.f@u.nus.edu, ahurriyetoglu@ku.edu.tr, t.caselli@rug.nl, nelleke.oostdijk@ru.nl,
nomoto@acm.org, hansi.hettiarachchi@mail.bcu.ac.uk, iqra@nlp.cic.ipn.mx, onuruca@mersin.edu.tr,
f.liza@uea.ac.uk, tianhu@ethz.ch

Abstract

Despite the importance of understanding causality, corpora addressing causal relations are limited. There is a discrepancy between existing annotation guidelines of event causality and conventional causality corpora that focus more on linguistics. Many guidelines restrict themselves to include only explicit relations or clause-based arguments. Therefore, we propose an annotation schema for event causality that addresses these concerns. We annotated 3,559 event sentences from protest event news with labels on whether it contains causal relations or not. Our corpus is known as the Causal News Corpus (CNC). A neural network built upon a state-of-the-art pre-trained language model performed well with 81.20% F1 score on test set, and 83.46% in 5-folds cross-validation. CNC is transferable across two external corpora: CausalTimeBank (CTB) and Penn Discourse Treebank (PDTB). Leveraging each of these external datasets for training, we achieved up to approximately 64% F1 on the CNC test set without additional fine-tuning. CNC also served as an effective training and pre-training dataset for the two external corpora. Lastly, we demonstrate the difficulty of our task to the layman in a crowd-sourced annotation exercise. Our annotated corpus is publicly available, providing a valuable resource for causal text mining researchers.

Keywords: causality, event causality, text mining, natural language understanding

1. Introduction

Causality is a core cognitive concept and appears in many natural language processing (NLP) works that aim to tackle inference and understanding (Jo et al., 2021; Dunietz et al., 2020; Feder et al., 2021). Generally, a causal relation is a semantic relationship between two arguments known as cause and effect, in which the occurrence of one (cause argument) causes the occurrence of the other (effect argument) (Barik et al., 2016).

Figure 1 depicts examples of sentences expressing causality, and ones that do not. Notice that causality can be expressed in various ways: The first causal example is signaled by the explicit causal marker *"due to"* while the second example has causality indicated by alternative lexicalizations such as *"created"*. In the last example, an implicit causal relation exists between the verbless clause and the matrix clause, conveying that the dissatisfaction with the package leads to the workers' sit-in. For sentences without causality, they must be missing either a cause or effect argument, or both.

Causal Question Answering and Generation applications (Dalal et al., 2021; Hassanzadeh et al., 2019; Stasaski et al., 2021) are some immediate downstream

Figure 1: Annotated examples from Causal News Corpus. Causes are in pink, Effects in green and Signals in yellow. Note that both Cause and Effect spans must be present within one and the same sentence for us to mark it as *Causal*.

Natural Language Understanding (NLU) applications popular in NLP today. Despite the importance of identifying causality in text, datasets are limited (Asghar, 2016; Xu et al., 2020; Tan et al., 2021; Yang et al.,

2021) and oftentimes, different researchers craft their datasets with different rules, leaving users with no proper way to compare models across datasets.

Our work is directed at annotating parts of the multilingual protest news detection dataset (Hürriyetoğlu et al., 2021a; Hürriyetoğlu et al., 2021b; Hürriyetoğlu, 2021) for event causality. Our contributions[1] are as follows:

- We created the Causal News Corpus (CNC), which is a corpus of event sentences annotated with binary labels indicating whether each sentence contains a causal relationship or not[2].

- We showed that a neural network built on a state-of-the-art pre-trained language model could predict the causality labels with 81.20% F1, 77.81% Accuracy, and 54.52% Matthews Correlation Coefficient on test set.

- We crafted annotation guidelines that align with existing event causality and linguistic schemes. Our experiments show that our dataset is indeed compatible with other causality corpora.

- We observed that the layman struggles to replicate the annotations by our experts in a crowd-sourced annotation exercise, demonstrating the difficulty of our task.

The rest of the paper is organized as follows: Section 2 reviews the literature on event causality. Section 3 describes the compilation and annotation of the corpus. Section 4 describes the baselines we created for causal event sentence classification and the experiments that were conducted on CNC and two other datasets. Section 5 outlines experiences from a crowd-sourcing annotation exercise. Finally, Section 6 concludes the paper and discusses potential avenues for future work.

2. Related Work

The extraction of causality from text remains a significant challenge because semantic understanding of the context and world knowledge is oftentimes needed. In computational linguistics, a lot of effort has been put into extracting causal knowledge from text automatically (Blanco et al., 2008; Do et al., 2011; Kontos and Sidiropoulou, 1991; Riaz and Girju, 2013).

Previous work has focused on event causality, creating corpora like CausalTimeBank (CTB) (Mirza et al., 2014) from news, CaTeRS (Mostafazadeh et al., 2016) from short stories and EventStoryLine (Caselli and Vossen, 2017) from online news articles. Such corpora include rich event features like event class attributes and temporal expressions in TimeML annotation format. However, these corpora are typically limited in size. For example, EventStoryLine (ESL) has 1,770 causal event pairs, CaTeRS has 488 causal links, while CTB has 318 causal event pairs. Our work contributes to this strand of literature by providing more data: We have identified 1,957 causal event sentences, each of which can contain multiple event pairs.

There is also a discrepancy between such event causality corpora and other causality corpora. Penn Discourse Treebank (PDTB) (Prasad et al., 2008; Webber et al., 2019; Prasad et al., 2006) is a corpus that annotates semantic relations (including causal relations) between clauses, expressed either explicitly or implicitly. This corpus is large (e.g. PDTB-3 has over $7,000$ examples for the CONTINGENCY.CAUSE sense alone) and is potentially useful for training an accurate event sentence classifier. Therefore, we believe that it will be beneficial to align the annotation guidelines of these two corpora types. We did so by constructing event causality annotation guidelines using linguistic rules adapted from PDTB-3. However, our corpus differs from PDTB-3 by focusing on event sentences and accepting more fine-grained arguments that do not necessarily form a clause. This approach of including more varied constructions of causality is similar to the work by BECauSE 2.0 (Dunietz et al., 2017).

When discussing causality in text, many corpora only focus on explicit relations (Girju and Moldovan, 2002; Dunietz et al., 2017). An explicit causal relation refers to an example where the cause-and-effect relation is expressed by means of a clear connective or some other causal signal. An example was presented earlier in the first sentence of Figure 1. Rule-based approaches (Girju and Moldovan, 2002; Khoo et al., 2000; Sakaji et al., 2008) have been shown to work well on such constructions. For example, researchers have used such approaches to semi-automatically curate datasets for cause-effect related tasks (Cao et al., 2016; Heindorf et al., 2020; Stasaski et al., 2021). However, in the real world, implicit relations are more common but more challenging to identify (Hidey and McKeown, 2016). Therefore, in our work, we do not limit ourselves to a pre-defined list of connectives but rather, include causal examples in more varied linguistic constructions for more realistic use cases. With respect to this, we differ from CTB, which only annotated explicit causal examples.

3. Compilation & Annotation

3.1. Data Source

CNC builds on the datasets featured in a series of workshops directed at mining socio-political events from news articles: Automated Extraction of Socio-political Events from News (AESPEN) in 2020 (Hürriyetoğlu et al., 2020b; Hürriyetoğlu et al., 2020a) and Challenges and Applications of Automated Extraction of

---

Socio-political Events from Text (CASE) in 2021 (Hür-riyetoğlu et al., 2021a; Hürriyetoğlu et al., 2021b; Hür-riyetoğlu, 2021). The data was based on randomly sampled articles from multiple sources and periods, and all annotations were performed by two annotators, adjudicated and spot-checked by a supervisor, and corrected further semi-automatically (Hürriyetoğlu et al., 2021c; Yörük et al., 2021). In total, 869 news documents and 3,559 English sentences were available. We annotated all 3,559 sentences in this paper. We utilized the dataset from CASE 2021 directly and some sentences were not well processed. We noticed that a few examples were actually comprised of more than one sentence and titles were still included in the texts. In our next data release, we will redo the parsing of these examples.

3.2. Guidelines

CNC includes annotations for binary causal event classification in sentences; We labeled sentences to be *Causal* or *Non-causal*. Sentences must contain at least a pair of events, defined as as things that happen or occur, or states that are valid, following TimeML (Saurı et al., 2006). For causality, we utilized the definition of CONTINGENCY from PDTB-3, which assigns this label for samples where "one argument provides the reason, explanation or justification for the situation described by the other" (Webber et al., 2019). The following four senses were recognized as *Causal*:

- CAUSE: Arg1 and Arg2 are causally influenced but not in a conditional relation. Example connectives are *"because"* and *"since"*.

- PURPOSE: One Arg describes the goal while the other Arg describes the action undertaken to achieve the goal. Example connectives are *"in order to"* and *"so that"*.

- CONDITION: One Arg presents an unrealized situation, that when realized, would lead to the situation described by the other Arg. Example connectives are *"if"* and *"as long as"*.

- NEGATIVE-CONDITION: One Arg presents an unrealized situation, that if it does not occur, would lead to the situation described by the other Arg. Example connectives are *"till"* and *"unless"*.

Since our work is event-based, the senses that provided reasons for the speaker to utter a speech, or the hearer to have a belief (i.e. +SPEECHACT or +BELIEF types) were treated as *Non-causal*.

A major difference between our approach and PDTB is that we permitted non-clausal elements such as phrases in our annotation scheme as arguments as long as these spans meet our definition of event. PDTB is a discourse bank that restricts its argument spans to discourse relations. As noted by Dunietz et al. (2017), causal spans of multiple linguistic forms are excluded from PDTB.

For instance, in PDTB, verb signals like *"caused"* are not annotated in the sentence *"Speculation about Coniston has caused the stock to rebound from a low of $145."*. Consequently, PDTB also did not annotate the spans before and after *"caused"* as Cause and Effect arguments. Given our focus on news events, more expressions of causality are required for proper interpretation. Therefore, on top of PDTB-approved clauses, we allowed for the following to be recognized as an argument: noun phrases (including nominalizations)[3], and verb phrases (including ones that fall outside of a coordinated structure, provided that the corresponding span of its relation is a noun phrase that would complete the verb phrase to form a clause). In general, these relaxations of rules allowed us to annotate more fine-grained arguments, including ones within clauses. To ground our annotations, we utilized the five tests for causality based on the work by Grivaz (2010) and Dunietz et al. (2017). In Table 1, we report some examples to illustrate the applications of the causality tests and the resulting classification labels. Additional details and examples are provided in Appendix A.

Five Tests for Causality

1. **Why:** The example is not causal if the reader is unable to construct a "Why" question regarding the Effect.

2. **Temporal order:** The example is not causal if the Cause does not precede the Effect in time.

3. **Counterfactual:** The example is not causal if the Effect is equally likely to occur or not occur without the Cause.

4. **Ontological asymmetry:** The example is not causal if the reader can readily swap the Cause and Effect claims in place.

5. **Linguistic:** The example is likely to be causal if it can be rephrased into "X causes Y" or "Due to X, Y."

CAUSE types Quoting the PDTB-3 annotation manual (Webber et al., 2019), "CAUSE.REASON (CAUSE.RESULT) is used when Arg2 (Arg1) gives the reason, explanation or justification, while Arg1 (Arg2) gives its effect". Within the CAUSE sense, we noticed that its sub-senses, CAUSE.REASON and CAUSE.RESULT, fit into the five tests framework well. In particular, we were able to construct a question like "Why did `<effect>`?", and answer with "Because (of) `<cause>`." in an extractive manner.

However, we required some relaxations in the construction of the "Why" test to be able to include other senses into our corpus. We highlight the relaxations as follows:

---

[3]Some examples of noun phrases describing event arguments: *"the fire"* and *"the protest in May"*.

Sentence	Causality Tests					Label
	Why?	Temporal Order	Counterfact.	Ontological Asymmetry	Linguistic	
The protests spread to 15 other towns and resulted in two deaths and the destruction of property .	✓	✓	✓	✓	✓	*Causal*
Chale was allegedly chased by a group of about 30 people and was hacked to death with pangas, axes and spears .	✗	✓	✗	✓	✗	*Non-causal*
The strike will continue till our demands are conceded .	✓	✓	✓	✓	✓	*Causal (Neg. Cond.)*

Table 1: Examples illustrating the applications of the Tests for Causality (Grivaz, 2010; Dunietz et al., 2017). Cause in pink, Effect in green; potential Cause in gray. Signals are not marked.

PURPOSE types We permitted the answer to include additional words like "In order to (achieve the goal of) <cause>". In this setting, the cause is a goal, or a justification for the effect action. See Appendix A Example (3).

CONDITION types We allowed the relaxation of the question to include modal terms, for example: "Why could/would <effect> occur?". See Appendix A Example (7).

CAUSE.NEGRESULT and NEGATIVE-CONDITION types Quoting the PDTB-3 annotation manual (Webber et al., 2019), "CAUSE.NEGRESULT is used when Arg1 gives the reason, explanation or justification that prevents the effect mentioned in Arg2". For CAUSE.NEGRESULT and NEGATIVE-CONDITION senses, we allowed a further relaxation of the question to include "prevent" or "cause-to-end" terms, as in: "Why was <effect> prevented?", "Why could be <effect> prevented?" or "Why did <effect> end?". See Appendix A Example (8).

For samples with multiple sentences, we annotated each sentence independently, that is, we do not consider cross-sentence causality. We acknowledge that by omitting cross-sentence examples, we potentially lose out between a 1/10 to a 1/5 of additional causal examples, based on proportions estimated from CTB and ESL respectively. Titles were also not annotated.

3.3. Workflow

Five annotators and one curator were involved in the annotation effort. Figure 2 summarizes the different annotation iterations to build CNC.

3.3.1. Initial Training

The initial training phase for annotators corresponds to Round 1 in Figure 2. Guidelines were presented at the beginning of the annotation phase, and examples outside of CNC dataset were ran through with annotators.
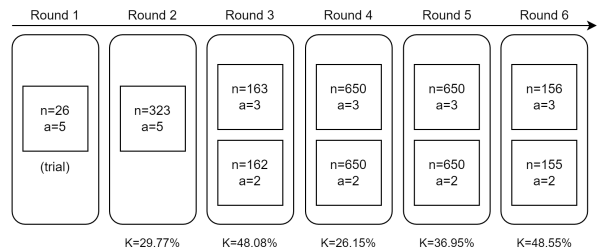

Figure 2: Summary of annotation workflow. n: sample size for the subset. a: number of annotators for the subset. κ: Krippendorff's Alpha score for samples per round. Round 1 was a trial training round where annotations were discarded. All subsequent rounds were for the training set except Round 6 which was for the test set.

We found it difficult to arrive at a consensus on the final label across annotators during this initial discussion. This experience motivated us to update our guidelines after each round of annotation to continually improve annotators' understanding of the guidelines.

3.3.2. Guidelines Refinement

The difficulty in achieving alignment across annotators, especially at the initial stages, arose from the following three issues: Firstly, since our task challenges literacy, comprehension, and cognitive understanding of a sentence, it is a complex and involved task. Thus, a significant amount of time is required for careful reading of the annotation manual and examples to improve familiarity. Secondly, sensitivity to causal cues and signals (including implicit signals) takes time to develop, and might not be apparent to some until later rounds. Thirdly, different annotators rely on different guidelines or feedback mechanisms to improve their accuracy to identify causal labels faithful to the guidelines.

Therefore, we were motivated to update our guidelines in an iterative fashion after each annotation round to refine the descriptions and include more examples that would help improve alignment across annotators. A group discussion always followed at the end of each annotation round for annotators and curator to share feedback and agree on the final guidelines. We also allowed annotators to skip examples if they were unable to come up with a suitable label.

For any major changes in guidelines that might affect previous rounds of annotations, the curator identified such examples and discussed with all annotators before the final annotation is amended. In total, 33 examples had their labels changed in this manner.

3.3.3. Final Annotation

In Round 2, we achieved a Krippendorff's Alpha (κ) score of 29.77% across five annotators. After the second annotation round, we split the five annotators into two groups to cover more data within our time frame. Concurrently, we also noticed an improvement in the agreement scores to 48.08%. By the third annotation round, the guidelines had no further changes. Surprisingly, we noticed agreement scores fell to 26.15% in Round 4. This could be due to the much larger dataset presented to annotators in this round, causing reduced focus. In fact, we noticed the main reason for the drop arose from one annotator in the first group. Excluding his contribution, the agreement increased to a reasonable score of 39.34%. After working out misunderstandings, the subsequent two rounds of annotations achieved consistent agreement scores of 36.95% and 48.55%. Across all rounds, our dataset has an agreement score of 34.99%.

In our final corpus, every example was annotated by at least two annotators. If only one annotation is present because the other annotator skipped that example, the curator's vote is final. The curator also resolved any ties present throughout the annotation exercise. In the end, we obtained 3,559 annotated sentences, with 1,957 marked as *Causal* and 1,602 marked as *Non-causal*. This is reflected in the first row of Table 2, alongside data sizes for two other external datasets we experimented with in the next section. We followed the train-test split of our data source, and also provide the breakdown of data sizes by class labels in Table 2.

	Causal	*Non-causal*	Total
CNC	1,957	1,602	3,559
CNC Training	1,781	1,467	3,248
CNC Test	176	135	311
PDTB-3	9,917	18,358	28,275
CTB	318	1,418	1,736

Table 2: Data sizes by class label.

4. Experiments

In this section, we present a battery of baseline models and results of causal sentence classification on the CNC data and the two existing news corpora, PDTB-3 (Webber et al., 2019) and CTB (Mirza et al., 2014).

4.1. Evaluation Metrics

For all experiments, we evaluated the predictions against true labels using five evaluation metrics: Precision (P), Recall (R), F1, Accuracy (Acc) and Matthews Correlation Coefficient (MCC) (Matthews, 1975). Formulas are provided below in Equations (1) to (5), where $TP/TN/FP/FN$ stands for True Positive/True Negative/False Positive/ False Negative respectively. All metrics were calculated using the `scikit-learn` (Pedregosa et al., 2011) package.

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

4.2. Models

4.2.1. Dummy Baselines

To create naive baselines, we obtained dummy predictions using two approaches. For the first row indicated by "All *Causal*" of Table 3, we predicted all examples to be *Causal*, which is the majority label of our dataset. Notice that the Recall score is perfect (100%), because all *Causal* examples will be identified if we predict all as *Causal*. However, Precision is low (56.59%), because all *Non-causal* examples were wrongly identified as *Causal*. In the subsequent discussions, we focus our key evaluation metrics to be F1, Accuracy and MCC. For the second row, we randomly assigned the *Causal* or *Non-causal* labels to each example based on the ground truth distribution from CNC's training set, averaged over $1,000$ runs.

4.2.2. Neural Network Baselines

We constructed a baseline model by fine-tuning the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019). In particular, we used `bert-base-cased` from Huggingface (Wolf et al., 2020) with AdamW Cross Entropy Loss. The learning rate was $5e - 05$ with linear scheduling. Train batch size was 32, and number of epochs was 10. The model pipeline is as follows: Tokenized sentences aree run through a BERT transformer module. The pooled output is obtained by taking the hidden state corresponding to the first token. Finally, a

#	Training Set	F1	P	R	Acc	MCC
1	All *Causal*	72.28	56.59	**100.00**	56.59	0.00
2	Random	55.72	56.61	54.92	50.66	0.00
3	CNC Training	**81.20**	78.01	84.66	**77.81**	**54.52**
4	PDTB-3	55.43	**81.32**	42.05	61.74	32.09
5	PDTB-3 Bal	64.45	77.60	55.11	65.59	34.75
6	CTB	27.36	80.56	16.48	50.48	17.49
7	CTB Bal	64.05	75.38	55.68	64.63	32.13

Table 3: Metrics from predictions on CNC Test Set using various settings. F1, Precision (P), Recall (R), Accuracy (Acc) and Matthews Correlation Coefficient (MCC) are reported in percentages (%). Highest score per column is indicated in boldface.

binary classification head is applied to obtain the predicted logits.

Additionally, we constructed a Long-Short Term Memory (LSTM) model (Hochreiter and Schmidhuber, 1997) with fastText embeddings (Bojanowski et al., 2017). The LSTM model performed on par with the BERT model for F1 but had lower MCC values than the BERT model. Overall, the BERT model outperformed the LSTM model. Details about the LSTM baseline and its evaluations are in Appendix B. Our subsequent analyses focuses on the BERT baseline.

4.3. Training on CNC

From the third row in Table 3, when we trained on our training set and applied on the test set, we achieved a reasonable F1 score of 81.20%, Accuracy of 77.81%, and MCC of 54.52%, each exceeding the dummy baselines scores reflected in the first two rows. This finding demonstrates that our annotations are internally consistent and reliable for a simple baseline model to learn effectively.

4.4. Training on External Corpora

In this subsection, we check the transferability of existing event causality and causality corpora on CNC. To conduct the experiment, we first trained on only the external corpus (i.e. PDTB-3 or CTB) and then applied the trained model to predict causal labels on CNC test set.

For PDTB-3, we trained on 28,275 examples (9,917 as *Causal* and 18,358 as *Non-causal*). We obtained this number of examples by selecting all causal examples based on the four senses described in Section 3.2. Next, we added the *Non-causal* examples and performed de-duplication such that if two examples from the same document overlap in some part but have a different causal label, we dropped the *Non-causal* example. Once the model is trained, we used it to predict the causal labels of CNC test set. According to Row 4 of Table 3, there is some overlap in the annotation schemes of PDTB and CNC. Therefore, the model was able to achieve Accuracy and MCC scores of 61.74% and 32.09%, which are higher than the first two dummy

baselines. However, the F1 score is only at 55.43%, comparable to the Random baseline and worse than the All-*Causal* assignment. The low F1 score is driven by the low Recall score, which indicates that a model trained on PDTB-3 is unable to detect most of the true *Causal* examples. However, its high Precision score indicates the model does return high-quality predictions. The low Recall could be explained by the fact that we included guidelines in our annotation schema that are different from PDTB's schema. For example, we were not strict about arguments having to be a proper clause. That is, for a sentence like *"John caused the fire and walked away."*, we would have marked it as *Causal*. However, since *"John"* and *"the fire"* are not clauses, PDTB would not mark this example as *Causal*. As a result, a model trained on PDTB would also recognize fewer causal constructions.

To tackle the issue of imbalanced class distribution in PDTB-3, which might affect model performance, we reran an experiment with randomly sampled *Non-causal* examples to match the number of *Causal* examples. This corpus is referred to as "PDTB-3 Bal" ($n = 19,834$) in Row 5 of Table 3. All metrics, except Precision, improved relative to the imbalanced PDTB-3 setting.

We repeated the experiment with examples obtained from CTB. For this corpus, we had 1,736 examples (318 as *Causal* and 1,418 as *Non-causal*) to train with. For sentences marked with "CLINK" in CTB, we denoted them as *Causal*. All remaining sentences were marked as *Non-causal*. Note a possible complication arises because while all *Non-causal* examples are single sentences, *Causal* examples might span across a few sentences. Performance was dismal, returning comparable or worse scores for F1 and Accuracy than both dummy baselines. We hypothesized that this poor outcome is a result of CTB's extremely imbalanced class distribution. Indeed, when we retrained on a balanced sample, "CTB Bal" ($n = 636$), we observed performance improvements across all metrics in the last row of Table 3. A reasonable F1 score of 64.05%, Accuracy of 64.63% and MCC of 32.13% was obtained, similar to or slightly better than PDTB's performance,

despite CTB Bal being much smaller in size.

To conclude, we trained models on external corpora and used them directly to make predictions on our proposed CNC corpus. This demonstrates the transferability of existing causality corpora on our corpus. Nonetheless, since both external corpora contain multi-sentence examples and differ slightly in annotation guidelines, some performance differences are expected.

4.5. Training & Testing on External Corpora

We continue to study the compatibility of CNC with the two external corpora. In this section, we study if a model trained on CNC will be helpful for event sentence causality prediction on unseen datasets like PDTB-3 and CTB.

For each of the three corpora, we trained on the whole corpus and applied the trained model on the remaining two corpora. For example, we trained on all examples from CNC (3,559 examples) and used the model to predict labels for PDTB-3 and CTB Bal. We experimented with the balanced corpus of CTB because the results of the experiments presented in Section 4.4 shown that the full CTB returns extremely poor performance when used for training. We also included the 5-Fold cross-validation (CV) scores as a reference: For each dataset, we randomly split the examples into five validation sets, and for each round, the corresponding remaining samples were used for training.

Table 4 reflects key experimental results. As expected, for every corpus, the CV training setup always returns the best performance. After which, for both PDTB-3 and CTB Bal, we noticed that a model trained on CNC always returns a higher performance across metrics. For example, when testing on CTB Bal, a model trained on PDTB-3 achieved 60.79% while a model trained on our corpus achieved 80.65% F1 score. We introduced a metric called Transferability Rate (TRF), which computes the mean score across test sets. TRF indicates how well a model trained on a given corpus works on average for unseen, external datasets. CNC obtained the highest TRF across the three corpora for both F1 and MCC, reflecting good transferability.

Our findings once again demonstrate that there exists some differences in the annotation schemes for causality, which reduces the transferability of trained models between corpora. Nevertheless, our corpus is the most transferable between the event causality (i.e. CTB) and linguistic causality (i.e. PDTB) corpora studied. Therefore, our annotation guidelines and dataset can serve as a bridge between these two strands of causality corpora.

4.6. Using CNC for Pre-Training

Since we believe CNC is transferable, we hypothesize that CNC will also be useful for training a pretrained model (PTM). Later, this PTM can be finetuned onto any other causality-based corpora. We term

the BERT PTM updated on all examples of CNC as 'CNC-PTM'. CNC-PTM was then fine-tuned on each external dataset (i.e. PDTB and CTB Bal). We again used the 5 CV setup[4], but reduced the training epochs to 2. We chose this low number in order to keep our model intact. The fine-tuning step only serves to adjust the model's alignment to better fit the new data distribution.

From Table 5, we indeed found improvements in the average performance across folds for both PDTB and CTB Bal. Logically, CNC-PTM is better initialized for causal sentence classification than the general `bert-base-cased` PTM for two reasons: (1) It had access to more causal examples and, (2) Its structure is more adapted for the classification task. Thus, performance improvements are expected. These findings again support our claim that CNC is generalizable to out-of-distribution datasets. Therefore, it was useful for updating a PTM.

5. Crowd-Sourcing Annotations

We experimented with crowd-sourced workers using Amazon Mechanical Turk (MTurk). We randomly selected 50 examples, of which 25 were *Causal* and 25 were *Non-causal*. For each causal label, we selected 13 examples that were more complex (with the percentage of the vote for the label being ≤ 0.75, while the remaining 12 examples were 'easier' examples with high agreements). The annotation manual was provided to the MTurk workers, and no restrictions were set on the type of workers we allowed for. We requested for five unique annotators to be assigned to each example. Therefore, 250 annotations were received[5], of which we had 44 unique workers.

Kappa scores were low at 1.62%. Evaluation metrics of these crowd-sourced annotations against our expert labels are reflected in Table 6. The second row reflects predictions taken based on the majority vote (i.e. mode out of five votes per example). The third row reflects outcomes when we compared each of the 250 annotations with our true labels independently. In both cases, performance of crowd-sourced workers is poor. Since most annotators labelled examples as *Causal*, the results are similar to the results from an All-*Causal* dummy baseline (shown in the first row). In fact, in terms of F1, Accuracy, and MCC, the workers' scores are close, if not worse, than the All *Causal* baseline. As highlighted in Section 3.3, the task is challenging and requires a thorough understanding of the annotation guideline to do well. Given that MTurk provides a low monitoring environment with no room for itera-

---

[4]For each external dataset and each fold, CNC-PTM was fine-tuned on the training set and applied onto the validation set. This means that for each external dataset, we would have 5 fine-tuned models at the end of the CV experiment.

[5]Although our dataset was balanced, crowd-sourced workers returned a skewed distribution in votes: 192 votes for *Causal* and 58 votes for *Non-causal*.

Training Set	Test Set							
	F1				MCC			
	CNC	PDTB-3	CTB Bal	TRF↑	CNC	PDTB-3	CTB Bal	TRF↑
CNC	**83.46**	58.38	80.65	**74.16**	**61.71**	30.68	59.11	**50.50**
PDTB-3	56.45	**74.45**	60.79	63.90	35.86	**61.36**	32.49	43.24
CTB Bal	59.10	49.21	**83.41**	63.90	32.10	17.48	**65.01**	38.20

Table 4: Metrics from predictions using different train and test sets. Diagonals per sub-table (i.e. when training set and test sets are the same) refers to 5-folds CV experiments. F1 and Matthews Correlation Coefficient (MCC) are reported in percentages (%). Transferability Rate (TRF) indicates how well a model trained on a given corpus works for unseen, external datasets: For example, the TRF for CNC is given by the average of its performance on PDTB-3 and CTB Bal, based on a particular metric (such as F1 or MCC). Highest score per column is indicated in boldface.

Dataset	PTM	F1	P	R	Acc	MCC
PDTB	`bert-base-cased`	74.45	**76.76**	72.31	82.60	61.36
	CNC-PTM	**75.19**	75.73	**74.69**	**82.71**	**61.95**
CTB Bal	`bert-base-cased`	83.41	77.25	**90.95**	81.91	65.01
	CNC-PTM	**84.68**	**80.14**	90.02	**83.80**	**68.30**

Table 5: Metrics from pre-trained model (PTM) experiments. F1, Precision (P), Recall (R), Accuracy (Acc) and Matthews Correlation Coefficient (MCC) are reported in percentages (%). Highest score per column per dataset is indicated in boldface.

	F1	P	R	Acc	MCC
All *Causal*	**66.00**	**50.00**	**100.00**	**48.40**	-3.89
Majority	61.97	47.83	88.00	46.00	-14.74
Each Vote	59.31	48.96	75.20	48.40	**-3.79**

Table 6: Metrics from crowd-sourced workers for a subset of 50 examples. F1, Precision (P), Recall (R), Accuracy (Acc) and Matthews Correlation Coefficient (MCC) are reported in percentages (%). Highest score per column is indicated in boldface.

tive feedback, it is understandable that crowd-sourced workers perform poorly (close to random assignment) at their tasks.

Nevertheless, this experiment provided a few learning points: Firstly, our manual annotations are not easy to identify by a layman, and skilled annotators are required. Similar to our annotation experience, it is necessary to have repeated discussions and feedback sessions with annotators so as to improve agreement scores. Secondly, compared to a BERT model trained on internal examples or external examples annotated with similar logic, the layman identifies causality poorly[6]. Caselli and Inel (2018) argues that discrepancies in the quality of crowd-sourced annotations from experts should be used as an estimate on how complex the task is. Together, these findings highlight that our corpus is a unique and valuable resource, requiring an immense amount of time and effort to create with experts.

6. Conclusion

Causality is an important cognitive concept that deserves to have a corpus dedicated to the definition and identification. CNC focuses on annotating causality in text, and so far, we have created an annotation schema to identify if event sentences contain causal relations or not.

CNC's annotation guidelines are constructed based on linguistic rules, and covers a wider array of causal linguistic constructions than previous works. We also demonstrated transferability between CNC and existing datasets that include causal relations like CTB and PDTB. Finally, the binary causal event classification task is challenging: layman workers achieved poor performance in our crowd-sourced annotation experiment. Therefore, CNC, which has been annotated by experts, is a valuable resource for researchers working in the causality and event causality space. CNC could potentially be relevant to NLU researchers working on downstream tasks too.

We are also organizing a shared task using CNC to promote the development of automatic causal text mining solutions[7]. In our next dataset release, we will clean some parts of our current dataset and perform additional curation of mispredicted examples in CV. We are also in the midst of adding fine-grained annotations, such as cause-effect-signal spans and causal concept labels.

---

[6]In terms beating dummy baselines, especially for Accuracy and MCC metrics.

[7]Our shared task competition page is at `https://codalab.lisn.upsaclay.fr/competitions/2299`.

7. Acknowledgements

8. Bibliographical References

Asghar, N. (2016). Automatic extraction of causal relations from natural language texts: a comprehensive survey. *arXiv preprint arXiv:1605.07895*.

Barik, B., Marsi, E., and Öztürk, P. (2016). Event causality extraction from natural science literature. *Res. Comput. Sci.*, 117:97–107.

Blanco, E., Castell, N., and Moldovan, D. (2008). Causal relation extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cao, M., Sun, X., and Zhuge, H. (2016). The role of cause-effect link within scientific paper. In Hai Zhuge et al., editors, *12th International Conference on Semantics, Knowledge and Grids, SKG 2016, Beijing, China, August 15-17, 2016*, pages 32–39. IEEE Computer Society.

Caselli, T. and Inel, O. (2018). Crowdsourcing StoryLines: Harnessing the crowd for causal relation annotation. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 44–54, Santa Fe, New Mexico, U.S.A, August. Association for Computational Linguistics.

Caselli, T. and Vossen, P. (2017). 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, August. Association for Computational Linguistics.

Dalal, D., Arcan, M., and Buitelaar, P. (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, June. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Do, Q., Chan, Y. S., and Roth, D. (2011). Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

Dunietz, J., Levin, L., and Carbonell, J. (2017). The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain, April. Association for Computational Linguistics.

Dunietz, J., Burnham, G., Bharadwaj, A., Rambow, O., Chu-Carroll, J., and Ferrucci, D. (2020). To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online, July. Association for Computational Linguistics.

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., et al. (2021). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*.

Girju, R. and Moldovan, D. I. (2002). Text mining for causal relations. In Susan M. Haller et al., editors, *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, May 14-16, 2002, Pensacola Beach, Florida, USA*, pages 360–364. AAAI Press.

Grivaz, C. (2010). Human judgements on causation in French texts. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Hassanzadeh, O., Bhattacharjya, D., Feblowitz, M., Srinivas, K., Perrone, M., Sohrabi, S., and Katz, M. (2019). Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5003–5009. International Joint Conferences on Artificial Intelligence Organization, 7.

Heindorf, S., Scholten, Y., Wachsmuth, H., Ngomo, A. N., and Potthast, M. (2020). Causenet: Towards a causality graph extracted from the web. In Mathieu d'Aquin, et al., editors, *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 3023–3030. ACM.

Hidey, C. and McKeown, K. (2016). Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany, August. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long

Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11.

Ali Hürriyetoğlu, et al., editors. (2020a). *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, Marseille, France, May. European Language Resources Association (ELRA).

Hürriyetoğlu, A., Zavarella, V., Tanev, H., Yörük, E., Safaya, A., and Mutlu, O. (2020b). Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 1–6, Marseille, France, May. European Language Resources Association (ELRA).

Hürriyetoğlu, A., Mutlu, O., Yörük, E., Liza, F. F., Kumar, R., and Ratan, S. (2021a). Multilingual protest news detection - shared task 1, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 79–91, Online, August. Association for Computational Linguistics.

Hürriyetoğlu, A., Tanev, H., Zavarella, V., Piskorski, J., Yeniterzi, R., Mutlu, O., Yuret, D., and Villavicencio, A. (2021b). Challenges and applications of automated extraction of socio-political events from text (CASE 2021): Workshop and shared task report. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 1–9, Online, August. Association for Computational Linguistics.

Hürriyetoğlu, A., Yörük, E., Mutlu, O., Duruşan, F., Yoltar, Ç., Yüret, D., and Gürel, B. (2021c). Cross-context news corpus for protest event-related knowledge base construction. *Data Intelligence*, 3(2):308–335, 06.

Ali Hürriyetoğlu, editor. (2021). *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, Online, August. Association for Computational Linguistics.

Jo, Y., Bang, S., Reed, C., and Hovy, E. H. (2021). Classifying argumentative relations using logical mechanisms and argumentation schemes. *Trans. Assoc. Comput. Linguistics*, 9:721–739.

Khoo, C. S. G., Chan, S., and Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 336–343, Hong Kong, October. Association for Computational Linguistics.

Kontos, J. and Sidiropoulou, M. (1991). On the acquisition of causal knowledge from scientific texts with attribute grammars. *International Journal of Applied Expert Systems*, 4(1):31–48.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Mirza, P., Sprugnoli, R., Tonelli, S., and Speranza, M. (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, April. Association for Computational Linguistics.

Mostafazadeh, N., Grealish, A., Chambers, N., Allen, J., and Vanderwende, L. (2016). CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California, June. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., and Webber, B. L. (2006). The penn discourse treebank 1.0 annotation manual. *IRCS Technical Reports Series*.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The penn discourse treebank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Riaz, M. and Girju, R. (2013). Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 21–30.

Sakaji, H., Sekine, S., and Masuyama, S. (2008). Extracting causal knowledge using clue phrases and syntactic patterns. In Takahira Yamaguchi, editor, *Practical Aspects of Knowledge Management, 7th International Conference, PAKM 2008, Yokohama, Japan, November 22-23, 2008. Proceedings*, volume 5345 of *Lecture Notes in Computer Science*, pages 111–122. Springer.

Saurı, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). Timeml annotation guidelines version 1.2. 1.

Stasaski, K., Rathod, M., Tu, T., Xiao, Y., and Hearst, M. A. (2021). Automatically generating cause-and-

effect questions from passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 158–170, Online, April. Association for Computational Linguistics.

Tan, F. A., Hazarika, D., Ng, S.-K., Poria, S., and Zimmermann, R. (2021). Causal augmentation for causal sentence classification. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 1–20, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2019). The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Xu, J., Zuo, W., Liang, S., and Zuo, X. (2020). A review of dataset and labeling methods for causality extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1519–1531, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Yang, J., Han, S. C., and Poon, J. (2021). A survey on extraction of causal relations from natural language text. *CoRR*, abs/2101.06426.

Yörük, E., Hürriyetoğlu, A., Duruşan, F., and Yoltar, Ç. (2021). Random sampling in corpus design: Cross-context generalizability in automated multicountry protest event collection. *American Behavioral Scientist*, 0(0):00027642211021630.

Appendix

A.   Annotated Examples

The following are some examples of the expected causal label for each sentence, along with justifications using the Five Tests highlighted in Section 3.2.

(1) His attackers allegedly drank his blood .

   - *Non-causal*: There is only one event in this sentence.

(2) The protests spread to 15 other towns and resulted in two deaths and the destruction of property .

   - *Causal*: This sentence contains causal events.

   1. **Why:** Why were there "two deaths and the destruction of property"? Because "the protests spread to 15 other towns".

   2. **Temporal order:** Protests must spread (Cause) before deaths and destruction (Effect) can occur.

   3. **Counterfactual:** Deaths and destruction (Effect) are unlikely to occur if the protests did not spread (NegCause).

   4. **Ontological asymmetry:** Using Cause as an Effect to construct a question: Why did "The protests spread to 15 other towns"? Because of deaths and destruction – does not answer the question.

   5. **Linguistic:** "The protests spread to 15 other towns" causes "two deaths and the destruction of property".

(3) The three-member Farlam Commission, chaired by retired judge Ian Farlam, was established by President Jacob Zuma to probe into the violence and the deaths of 44 people in wage-related protests.

   - *Causal*: This sentence contains causal event of the Purpose sense.

   1. **Why:** Why was "the three-member Farlam Commision ... established by President Jacob Zuma"? In order "to probe into the violence and the deaths of 44 people in wage-related protests".

   2. **Temporal order:** The probe purpose (Cause) must occur before establishing the commission (Effect).

   3. **Counterfactual:** Establishing the commission (Effect) is unlikely to occur if the President did not have the goal to probe into the protests (NegCause).

4. **Ontological asymmetry:** Using Cause as an Effect to construct a question: Why did the President "probe into violence and deaths of people in wage-related protests"? Because the Commission was established – does not answer the question.

5. **Linguistic:** "to probe into the violence and the deaths of 44 people in wage-related protests" causes "The three-member Farlam Commission , chaired by retired judge Ian Farlam , was established by President Jacob Zuma".

(4) Chale was allegedly chased by a group of about 30 people and was hacked to death with pangas, axes and spears.

- *Non-causal*: This sentence exhibits temporal events that have no causal connections. We highlight the tests it fails below:

1. **Why:** Why was Chale "hacked to death with pangas, axes and spears"? Because "Chale was allegedly chased by a group of about 30 people". – This test fails since the Cause does not really explain the reason for why there was the Effect.

2. **Counterfactual:** The counterfactual is "The hacking (Effect) is unlikely to occur if the chasing did not take place (NegCause)." – This test fails since hacking can occur without chasing. In fact, without chasing, Chale might have been hacked earlier.

(5) "I observed the attack on the police, I have no doubt about it," Modiba said during cross-examination.

- *Causal*: This sentence contains a description of an implicit causal event relation within a speech.

1. **Why:** Why did Modiba "have no doubt about it"? Because Modiba "observed the attack on the police".

2. **Temporal order:** Modiba must observe (Cause) before claiming he had no doubts (Effect).

3. **Counterfactual:** Modiba cannot claim that he has no doubts (Effect) if he had not observed the attack (NegCause).

4. **Ontological asymmetry:** Using Cause as an Effect to construct a question: Why was it that Modiba "observed the attack on the police"? Because he had no doubt about it – does not answer the question.

5. **Linguistic:** "I observed the attack on the police" causes "I have no doubt about it".

(6) Both Anoop and Ramanan are also accused in the case related to attack on the Nitta Gelatin office at Panampilly Nagar last year.

- *Non-causal*: The inclusion of "also" in "also accused" renders that when constructing an answer for the Effect, we need to explain why they are accused in addition to someone else that has already been accused. Therefore, when performing the "Why" test, the answer seems incomplete. Hence, we do not consider this sentence to be causal.

1. **Why:** Why were "both Anoop and Ramanan also accused in the case"? Because of the "attack on the Nitta Gelatin office at Panampilly Nagar last year ." – This does not answer the question.

2. **Linguistic:** "attack on the Nitta Gelatin office at Panampilly Nagar last year " cause "Both Anoop and Ramanan are also accused in the case". – This does not work.

(7) Striking mineworkers have threatened to halt all mining operations if their employers do not accede to their pay demand.

- *Causal*: This sentence contains CONDITION relations, which we interpret as causal.

1. **Why:** Why could "halt all mining operations" occur? Because "their employers do not accede to their pay demand".

2. **Temporal order:** Their pay demands must be acceded (Cause) before the halting of mining operations (Effect).

3. **Counterfactual:** The halting of mining operations (Effect) is unlikely to occur if their pay demands were acceded (NegCause).

4. **Ontological asymmetry:** Using Cause as an Effect to construct a question: Why was it that "their employers do not accede to their pay demands"? We cannot find an answer span for this question in the original Effect span.

5. **Linguistic:** "their employers do not accede to their pay demand" causes "halt all mining operations".

(8) The strike will continue till our demands are conceded.

- *Causal*: This sentence contains NEGATIVE-CONDITION relations, which we interpret as preventive causal type.

1. **Why:** Why could the event that "the strike will continue" be prevented? Because "our demands are conceded".

Setting	F1	P	R	Acc	MCC
All *Causal*	72.28	56.59	**100.00**	56.59	0.00
Random	55.72	56.61	54.92	50.66	0.00
CNC Training	**78.22**	72.68	84.66	**73.31**	**45.15**
PDTB-3	56.68	66.41	49.43	57.23	16.90
PDTB-3 Bal	68.75	56.62	87.50	54.98	0.14
CTB	33.79	**86.05**	21.02	53.38	23.80
CTB Bal	65.23	71.14	60.23	63.67	28.15

Table 7: Metrics from predictions with the LSTM model on CNC Test Set using different training sets. F1, Precision (P), Recall (R), Accuracy (Acc) and Matthews Correlation Coefficient (MCC) are reported in percentages (%). Highest score per result column is indicated in boldface.

2. **Temporal order:** The demands must be conceded (Cause) before preventing the strike from continuing (NegEffect).

3. **Counterfactual:** The strike cannot be stopped from continuing (NegEffect) if the demands are not conceded (NegCause).

4. **Ontological asymmetry:** Using Cause as an Effect to construct a question: Why was it that "our demands are conceded"? Because "the strike could be stopped" – does not answer the question.

5. **Linguistic:** "Our demands are conceded" prevents "the strike will continue".

B.   Other Baseline Models

As described in Section 4, in addition to the BERT model, we constructed a LSTM model as one of our baselines, considering its ability to account for long-distance relations between words. In this model, firstly, an embedding layer initialized with fastText's Common Crawl 300-dimensional embeddings (Bojanowski et al., 2017). We experimented with GloVe 300-dimensional embeddings (Pennington et al., 2014) and a concatenation of GloVe and fastText in our initial experiments and selected fastText because it outperformed others. This layer is followed by two bi-directional LSTM layers with a dense layer on top. The final prediction is obtained via a softmax layer. For our training parameters, we used a maximum sequence length of 128, batch size of 32, a learning rate of $1e^{-3}$ with Adam optimizer and epochs of 20 with early stopping patience of 5. The obtained results are summarized in Table 7. We observed that performance of this system is significantly lower than BERT for all scenarios we tested. However, conclusions regarding the transferability of PDTB-3 and CTB Bal onto CNC (based on F1 and MCC scores) holds.