BoschAI @ Causal News Corpus 2023: Robust Cause-Effect Span Extraction using Multi-Layer Sequence Tagging and Data Augmentation

Timo Pierre Schrader[1,2] **Simon Razniewski**[1] **Lukas Lange**[1] **Annemarie Friedrich**[2]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]University of Augsburg, Germany
timo.schrader|simon.razniewski|lukas.lange@de.bosch.com
annemarie.friedrich@informatik.uni-augsburg.de

Abstract

Understanding causality is a core aspect of intelligence. The Event Causality Identification with Causal News Corpus Shared Task addresses two aspects of this challenge: Subtask 1 aims at detecting causal relationships in texts, and Subtask 2 requires identifying signal words and the spans that refer to the cause or effect, respectively. Our system, which is based on pre-trained transformers, stacked sequence tagging, and synthetic data augmentation, ranks third in Subtask 1 and wins Subtask 2 with an F1 score of 72.8, corresponding to a margin of 13 pp. to the second-best system.

1 Introduction

In this paper, we describe our approach to the Event Causality Identification with Causal News Corpus shared task (Tan et al., 2023), which took place at The 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023). The task, which builds on the 2022 iteration of the same shared task (Tan et al., 2022a), but including more labeled data, targets the detection and extraction of causal relationships. In Subtask 1, participating systems need to decide whether a sentence contains any causal relationship. Subtask 2 requires extracting the spans that denote cause, effect, and trigger words (if any).

Our system leverages pre-trained transformer encoders and synthetic data augmentation methods, and ranks third in Subtask 1. We address Subtask 2 using a supervised sequence labeling model, which wins by a margin of 13 percentage points in terms of F1 over the second-best system. We model multiple causal chains per sentence via stacked labels and find that synthetic data augmentation consistently improves performance. Our code is publicly available.[1]

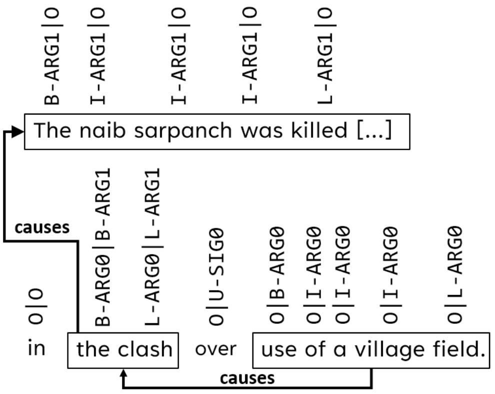

Figure 1: Our proposed modeling technique for extracting causal relationships (Subtask 2) using stacked BILOU labels. `ARG0` = cause, `ARG1` = effect.

2 Dataset and Task

The Causal News Corpus (CNC, Tan et al., 2022b) consists of 3767 sentences extracted from news articles. CNC provides annotations of semantic relations of the form "*X* causes *Y*" that indicate a *causal* relationship between arguments *X* and *Y*. The definition of causality follows that of the CONTINGENCY label in the PDTB-3 corpus (Webber et al., 2019), which is used when a statement provides the reason, explanation, or justification for another event. Following TimeML (Pustejovsky et al., 2003), the definition of events includes both actions that happen or occur and states. As illustrated by the example in Figure 1, one event is the immediate effect of another, e.g., the event expressed by "the use of a village field" is the cause of that expressed by "the clash."

While following the definition of causal relations of PDTB-3, which focuses on causal relations between sentences or clauses, CNC provides span annotations for causes (`ARG0`), effects (`ARG1`) and signals (`SIG0`) within sentences. Spans may comprise one to several words. Their boundaries are

[1] https://github.com/boschresearch/boschai-cnc-shared-task-ranlp2023

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 38–43, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_005

not restricted to clause or constituent boundaries. Signals are expressions such as "has led to" or "causing," but not every causal relation annotation requires a signal. Of all annotated relations, 30% do not contain a signal, for example: "[Dissatisfied with the package$_{\text{Cause}}$], [workers staged an all-night sit-in$_{\text{Effect}}$]." The average signal length is 1.46 words. Tan et al. (2022b) describe the annotation guidelines in detail.

The shared task is divided into two subtasks: Subtask 1 is a binary classification problem, deciding whether a sentence contains a cause-effect chain or not. Subtask 2 deals with the more challenging problem of extracting the correct spans of cause, effect, and signal, where a sentence may contain more than one causal relation. In CNC, the maximum number of causal relations per sentence is four. Spans are annotated using XML-like tags: ⟨ARG0⟩ refers to causes, ⟨ARG1⟩ to effects, and ⟨SIG0⟩ to signals.

3 Modeling and Augmentation

In this section, we describe the neural architectures that we use to solve the two subtasks. To produce contextualized embeddings of the input sentences, we use BERT-Large (Devlin et al., 2019) and RoBERTa-Large (Liu et al., 2019).

3.1 Subtask 1

We implement a binary classifier to detect whether a sentence contains a cause-effect relation. The sentence-level [CLS] embedding is fed into a linear output layer that outputs a prediction on whether a sentence contains a cause-effect meaning or not. We design the output layer to yield two prediction scores, one for each class. During our experiments, we observe that the classifier has shown prediction bias towards negative samples. Hence, we apply a weighted cross entropy loss that upweights the positive samples.

3.2 Subtask 2

We model the problem of detecting cause, effect, and signal spans, potentially with multiple causal relations within a single sentence, as sequence tagging task using the BILOU labeling scheme (Alex et al., 2007). The BILOU scheme extends the commonly used BIO scheme by introducing two additional markers, where "L" denotes the end of a multi-token sequence and "U" refers to a single-token entity. For example, the argument span "Beijing launched a campaign" has the label sequence [B-ARG1 I-ARG1 I-ARG1 L-ARG1] (ignoring BERT-specific subword tokens here). A linear layer on top of the embedding model produces the logits for all BILOU tags for each token individually. These logits are fed into a conditional random field (CRF, Lafferty et al., 2001) output layer, which computes the most likely consistent tag sequence.

However, this approach can only predict a single output sequence per sample, i.e., is not able to detect multiple causal chains in an instance. Consider the example shown in Figure 1. The expression "the clash" can be either the cause of one killing and 17 injuries or the effect of not being able to agree about the usage of a village field. As a result, there are two causal relations within this instance. To address this, we "stack" the BILOU labels by concatenating them using a pipe ("|") operator, similar to Straková et al. (2019), who also use a label stacking approach. As shown in Figure 1, this means that the word "clash" is tagged with L-ARG0|L-ARG1|O, which decodes to being the end of a cause in the first layer, being the end of an effect in the second one and not being part of any span in the third one.

To keep the label space manageable, we model three layers. There are only nine samples in the training set with four possible sequences. Without filtering, we would end up with about 39,000 labels. We only add stacked labels that occur in the training and validation data, resulting in roughly 300 three-layer BILOU labels. During evaluation, these stacked labels are split into their three distinct layers and each instance is evaluated separately. As a result, the model is able to predict up to three different causal relations per sentence.

3.3 Data Augmentation and Resampling

As for both subtasks, there is only limited training data available, we incorporate additional synthetic data into the training. In the 2022 edition of the shared task, several teams also experimented with data augmentation methods. Chen et al. (2022) trained BART (Lewis et al., 2020) to rephrase instances in the dataset. Kim et al. (2022) create additional data by adding the SemEval-2010 dataset (Hendrickx et al., 2010) and replacing words by their POS tag.

Augmenting using EDA Our first augmentation approach makes use of the Easy Data Augmenta-

Original Sentence	EDA Augmented Sentence
His arrest has sparked widespread protests by students, teachers as well as opposition parties.	His arrest has sparked widespread resist by student, teacher as advantageously as confrontation parties.
Month-long escalating protests to mark 4th anniversary of Mullivaikkal pogrom.	Month-long step up protests to mark off quaternary day of remembrance of Mullivaikkal pogrom.
They also rubbished suggestions that the student protests were losing steam [...]	They besides rubbish suggestions that the scholar protests were lose steam [...]

Table 1: Comparison between original sentences and their EDA-augmented counterparts. Differences are underlined.

tion (EDA, Wei and Zou, 2019) tool to generate additional training data for both subtasks. EDA offers different augmentation techniques: synonym replacement (*sr*), random word insertion (*ri*), random word deletion (*rd*), and random word swaps (*rs*). The percentage of words on which these techniques are applied are defined by hyperparameters α_{sr}, α_{ri}, α_{rd}, and α_{rs}.

For Subtask 1, we employ synonym replacement, random word insertion, and random swaps and generate four synthetic samples per original instance in the training set. This results in a training set five times as large as the original dataset with a total sample count of over 15.000 samples.[2] In Subtask 2, keeping the ordering of ⟨ARG0⟩, ⟨ARG1⟩ and ⟨SIG0⟩ consistent is of high importance. To avoid adding destructive noise to the training data, we only use synonym replacement and random insertion for this subtask. We add one augmented sample per single-relation instance, i.e., we do not augment data based on samples with more than one causal relation. We discard augmented samples that are invalid w.r.t. the annotation scheme. Data augmentation for the challenging multi-relation cases is an interesting direction for future research. The augmented training set contains 4.611 instances, i.e., about 1.500 more than the original set.

Table 1 shows three instances and their augmented counterparts. The first example shows a replacement of *opposition* by *confrontation*, which is not fully synonymous, but still related. In the second one, there is a synonym replacement of *4th* by *quaternary*. In the third example, noise is added by replacing “losing” with “lose”, illustrating that the data augmentation method does not control for grammatical correctness.

Oversampling of Multi-Relation Samples About 32% of all instances with at least one causal relation in the training set are labeled with more than one causal relation. Out of these, we sample 400 instances (with replacement) and add them to the training dataset. In contrast to EDA, we only use this setting only for Subtask 2.

Generating Samples using ChatGPT We experiment with GPT-3.5-turbo and prompt it to generate 100 novel samples containing causal relations that are similar to those of the CNC corpus. We prompt ChatGPT with multiple samples of the CNC train set, and the rules of placing ⟨ARG0⟩, ⟨ARG1⟩, and ⟨SIG0⟩, and let it generate novel samples. This additional data is only used for Subtask 2.

The ChatGPT-based data augmentation approach generates relatively simple examples by always sticking to a Cause-Signal-Effect or Effect-Signal-Cause structure without overlapping spans. Examples include “[The lack of rain$_{\text{Cause}}$] [caused$_{\text{Signal}}$] [the crops to fail and farmers to suffer losses$_{\text{Effect}}$].” and “[A decrease in greenhouse gas emissions$_{\text{Effect}}$] [was a result of$_{\text{Signal}}$] [the decrease in demand for fossil fuels$_{\text{Cause}}$]”.

4 Experimental Evaluation

This section describes our experimental results for both subtasks. Evaluation of Subtask 2 is performed using FairEval[3], which implements a relaxation of traditional hard-matching span evaluation metrics on sentences marked as containing a causal relation in the gold standard only. We train our on all samples of the train split, including those without causal relations.

4.1 Hyperparameters

To find the best learning rates and augmentation parameter combinations, we employ a grid search

[2] We noticed that the tool also clones each original sample in our implementation.

[3] https://huggingface.co/spaces/hpi-dhc/FairEval/tree/main

	Team	Precision	Recall	F1
1	DeepBlueAI	**83.2**	86.1	**84.7**
2	InterosML	81.6	87.3	84.4
3	BoschAI	80.0	87.9	83.8
	baseline	75.9	**89.2**	81.9

Table 2: Subtask 1: results on **test** of the best three systems and the baseline provided by Tan et al. (2023). Scores are based on the public leaderboard.

LM	Precision	Recall	F1	Accuracy
BERT	86.9	**89.7**	**88.3**	87.1
RoBERTa	**88.6**	88.1	**88.3**	**87.4**

Table 3: Subtask 1: results on **dev** (large model variants).

		All relations			Multi-relation		
		P	R	F1	P	R	F1
1	BoschAI	**84.4**	**64.0**	**72.8**	82.6	53.5	64.9
	- Cause	85.3	59.7	70.2	82.5	47.4	60.2
	- Effect	82.8	62.9	71.5	80.3	50.4	61.9
	- Signal	85.4	70.4	77.2	82.6	53.5	64.9
2	tanfiona*	60.3	59.2	59.7	-	-	-
3	CSECU-DSG	40.0	36.1	38.0	-	-	-

Table 4: Per-class scores on the **test** for Subtask 2 of our best scoring model using RoBERTa-Large and EDA. The last two rows show the results of the second- and third-best system. *System of Chen et al. (2022).

and refine the learning rates after an initial coarse-grained search ranging from $1e^{-7}$ to $9e^{-4}$ for the pre-trained language model. The binary classifier for Subtask 1 is trained with a learning rate of 8e-6, using the EDA augmented training data and a batch size of 32. For Subtask 1, we use the following parameter values for the different EDA techniques: $\alpha_{sr} = 0.4$, $\alpha_{ri} = 0.1$, and $\alpha_{rs} = 0.6$. We use a weighted cross entropy loss for this subtask, using a weight of 1.5 for class *causal*. For Subtask 2, we apply the following settings: $\alpha_{sr} = 0.4$ and $\alpha_{ri} = 0.5$.

The CRF-based tagger for Subtask 2 uses a learning rate of $7e^{-5}$ for the language model and the linear layer, whereas a learning rate of $3e^{-4}$ is applied on the CRF. During fine-tuning, EDA-augmentated data is included in the training set. Training the models is performed on Nvidia A100 GPUs using one GPU per run, which takes several hours per model. Early stopping is applied using the F1 score on the dev set and a patience of three epochs to select the best model. The models are optimized using AdamW (Loshchilov and Hutter, 2019) and an inverse square-root learning rate scheduler taken from Grünewald et al. (2021).

4.2 Results

In the following, we refer to the public leaderboard of the Event Causality Identification with Causal News Corpus shared task.[4] We report results on test as provided by the leaderboard evaluation script.

Subtask 1 Our RoBERTa-based binary classifier ranks third of 10 participants. Results are shown in Table 2, including the best two systems and the baseline by Tan et al. (2023). Among the top three, we achieve the best recall score. Qualitatively, we find that neither sentence length nor the presence of signal words are strongly correlated with misclassifications.

We report the results of our classifier that uses BERT-Large in comparison to RoBERTa-Large in Table 3 on the dev set (since we do not have access to the gold standard of test). Both models perform almost equally on this task, with RoBERTa outperforming BERT by a slight margin in terms of accuracy with a difference 0.3% pp.

Subtask 2 On this task, we compare our models against the baseline provided by Tan et al. (2023), which is the best performing system from the previous iteration of the shared task by team “1Cademy” (Chen et al., 2022). They also build upon a BERT-based embedding model, but output prediction scores for begin and end tokens of the respective spans. In order to produce consistent output, i.e., non-overlapping cause and effect spans and correctly ordered spans, they implement a beam-search algorithm on top that aims to find the top m most likely spans for each of the three types.

Per-label scores of our best-performing model and those of the other two competitors are shown in Table 4. Our best system is based on RoBERTa-Large with a CRF layer on top and trained on EDA-augmented data. Our system clearly outperforms the last year’s winning system by more than 13 percentage points in terms of F1 on the latest CNC data, exceeding precision by 24 percentage points. Our system performs best on the signal label, which could be explained by two factors: signals are much more repetitive in the corpus (with “to” occurring 293 times in the train data) and the average length

[4] https://codalab.lisn.upsaclay.fr/competitions/11784#results

	Cause			Effect			Signal			*avg*		
LM	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERT-Large	82.4	59.9	69.4	83.2	58.7	68.9	86.3	72.0	78.5	83.8	62.6	71.6
RoBERTa-Large	86.8	66.1	75.1	85.2	**68.5**	76.0	82.8	75.4	78.9	85.1	69.3	76.4
+ EDA*	86.4	**67.9**	76.1	**88.5**	67.9	**76.8**	85.0	**77.5**	**81.1**	86.8	**70.3**	**77.7**
+ Oversampling	87.5	67.7	**76.3**	86.8	66.2	75.1	**85.1**	74.3	79.3	86.6	68.8	76.7
+ ChatGPT	**88.4**	65.7	75.4	87.5	66.8	75.8	84.3	75.2	79.5	**86.9**	68.5	76.6

Table 5: Subtask 2 results on **dev**: precision, recall and F1 scores for cause, effect and signal span predictions. *Our system used to produce leaderboard scores.

Relations/Sentence	**Cause**	**Effect**	**Signal**
1	85.5	80.9	84.7
2	67.7	76.1	84.0
3	52.8	61.8	57.9

Table 6: Per-class F1 scores by the numbers of causal relations per sentence on **dev** for Subtask 2.

of 1.46 words is much smaller than those of causes (11.74) and effects (10.74). Table 4 also lists the results for multi-relation instances only, showing that recall drops for those instances.

Table 5 compares several settings, including various data augmentation techniques, by label on the dev set. We evaluate on the dev set because we do not have access to the gold standard of the test set.

First of all, using RoBERTa over BERT improves the average F1 score by 4.8 points in terms of F1. Next, all three data augmentation methods contribute performance improvements over the RoBERTa baseline with the recall of **Effect** being the only exception. Best overall results are achieved using EDA augmentation. However, ChatGPT-augmented significantly improves precision of **Cause** (1.6 points F1 over baseline) and also yields the best average precision. Tan et al. (2022b) also experiment with using two additional corpora, however, they do not get significant improvements, likely due to more different foci of the datasets. The synthetic data augmentation methods that we used have the advantage of producing training data very similar to CNC.

Finally, Table 6 breaks down results on dev split by single-relation, two-relation and three-relation instances. While scores for Effect and Signal remain high for two-relation instances, performance is much smaller (yet still strong) for three-relation instances.

5 Conclusion and Outlook

In this paper, we have described our modeling approach to the "Event Causality Identification with Causal News Corpus" shared task (CASE 2023). We have proposed a multi-layer sequence tagging model that aims at identifying causal relations within news-related sentences. Our approach significantly outperforms all participating systems in Subtask 2. Furthermore, we have shown that synthetic data augmentation methods are beneficial for this task. Our results indicate that careful modeling, more advanced data augmentation, and leveraging larger language models may be fruitful directions for further improvements.

References

Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Biological, translational, and clinical language processing*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics.

Xingran Chen, Ge Zhang, Adam Nik, Mingyu Li, and Jie Fu. 2022. 1Cademy @ causal news corpus 2022: Enhance causal span detection via beam-search-based position selector. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 100–105, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan Grünewald, Annemarie Friedrich, and Jonas Kuhn. 2021. Applying occam's razor to transformer-based dependency parsing: What works, what doesn't, and what is really necessary. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 131–144, Online. Association for Computational Linguistics.

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.

Juhyeon Kim, Yesong Choe, and Sanghack Lee. 2022. SNU-causality lab @ causal news corpus 2022: Detecting causality by data augmentation via part-of-speech tagging. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 44–49, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Rob Gaizauskas, Andrea Setzer, and Graham Katz. 2003. Timeml: A specification language for temporal and event expressions.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Tommaso Caselli, Onur Uca, Farhana Ferdousi Liza, and Nelleke Oostdijk. 2022a. Event causality identification with causal news corpus - shared task 3, CASE 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 195–208, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Onur Uca, Farhana Ferdousi Liza, and Nelleke Oostdijk. 2023. Event causality identification with causal news corpus - shared task 3, CASE 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*. Association for Computational Linguistics.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022b. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.