InterosML @ Causal News Corpus 2023: Understanding Causal Relationships: Supervised Contrastive Learning for Event Classification

Rajat Patel
Interos Inc. Arlington VA 22203, USA
rpatel@interos.ai

Abstract

Causal events play a crucial role in explaining the intricate relationships between the causes and effects of events. However, comprehending causal events within discourse, text, or speech poses significant semantic challenges. We propose a contrastive learning-based method in this submission to the Causal News Corpus - Event Causality Shared Task 2023, with a specific focus on Subtask 1 centered on causal event classification. In our approach we pre-train our base model using Supervised Contrastive (SuperCon) learning. Subsequently, we fine-tune the pre-trained model for the specific task of causal event classification. Our experimentation demonstrates the effectiveness of our method, achieving a competitive performance, and securing the 2nd position on the leaderboard with an F1-Score of 84.36.

1 Introduction

Understanding the intricate relationships between cause and effect within events is a fundamental aspect of language comprehension. Causal events, which provide insights into these connections, present semantic challenges when it comes to their classification and analysis in discourse, text, or speech.

We tackle the specific problem of causal event classification in Subtask 1 of the Causal News Corpus -Event Causality Shared Task 2023 (Tan et al., 2023) in our submission. This task involves accurately identifying and categorizing causal events, which plays a vital role in unraveling the underlying mechanisms behind real-world phenomena. Successful classification enables a wide range of applications, such as information extraction, summarization, and knowledge graph construction. To address this challenge, we propose an innovative approach that leverages SuperCon learning and source-aware sampling.

Contrastive learning has shown promising results in computer vision to learn a better and robust visual representations (Chen et al., 2020) and various natural language processing task like knowledge graph embeddings (Luo et al., 2021), text classification (Chen et al., 2022), entity linking (Yuan et al., 2022) and entity resolution (Brinkmann et al., 2023) etc. It allows the models to learn by contrasting positive and negative pairs, capturing informative representations.

The use of contrastive learning in text classification has been investigated in various contexts. For instance, the study by (Zuo et al., 2021) employed self-supervised learning techniques to address event causality identification in scenarios with limited annotated datasets. Similarly, (Chen et al., 2022) took an approach to incorporate contrastive learning with synthesized counterfactuals for data augmentation, demonstrating notable improvements in aspects such as counterfactual robustness, cross-domain generalization.

In this paper we apply the idea of SuperCon learning introduced by (Khosla et al., 2020) to the causal event classification task. Further, we loosely connect to the idea of source-aware sampling strategy introduced by (Peeters and Bizer, 2022) and modify it to suite the classification SubTask for pre-training the base encoder architecture.

Our methodology involves pre-training a transformer based encoder model using SuperCon Loss with naive source-aware sampling, followed by fine-tuning the pre-trained model on the causal event classification task. Through extensive experimentation and evaluation on the Causal News Corpus dataset, we demonstrate the effectiveness of our approach.

This paper's contributions can be summarized as follows: (1) Introducing contrastive learning as a method for causal event classification. (2) Achieving competitive performance in the Causal News Corpus - Event Causality Shared Task 2023,

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 60–65, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_008

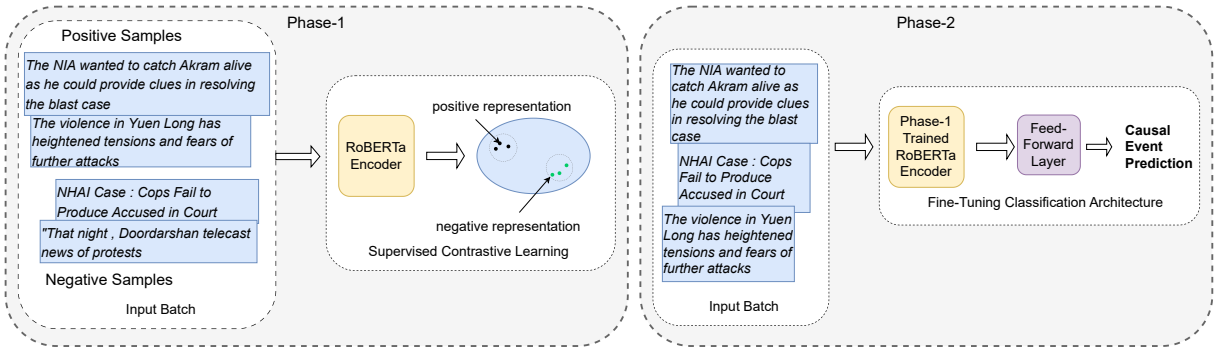

Figure 1: **Learning Phases for Causal Event Classification**: Phase-1: Pre-training with SuperCon | Phase-2: Fine-tuning for Causal Events

with the F1 Score of 84.36.[1]

2 Methodology

In this section, we present the methodology employed to address the causal event classification task. Our approach utilizes contrastive learning and consists of two main phases (Figure 1): (1) Pre-training the baseline transformer architecture with SuperCon, and (2) Fine-tuning the pre-trained model on the downstream classification task. For the encoder architecture, we adopt the RoBERTa base model[2] which has been shown to achieve strong results across different benchmark tasks (Liu et al., 2019).

2.1 Contrastive Pre-training

During the pre-training phase, we employ a batch creation process similar to the work of (Khosla et al., 2020) and augment it with the‘ naive source-aware sampling strategy introduced by (Peeters and Bizer, 2022). To train the encoder model, we create two copies of the input dataset. From the first dataset, we randomly select N records of input text x and subsequently sample another set of N records of input text x' from the second dataset, where we record in the batch (of size $2N$) has at least one corresponding record with the same label (even if it is a duplicate record only)

The RoBERTa encoder maps each input causal text record x to an embedding z as

$$z = \text{RoBERTa}(x). \tag{1}$$

To enhance the robustness and generalization of the record embeddings, we perform mean pooling on the encoder's output embeddings

$$z = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{2}$$

and normalize them using the L^2-norm

$$z \rightarrow \frac{z}{\|z\|} \tag{3}$$

— a strategy effectively employed by (Brinkmann et al., 2023) for entity resolution tasks. To train the parameter of the encoder RoBERTa architecture we apply SupCon Loss to to cluster or position records with the same label more densely within the embedding space.

The SuperCon Loss employs the principle of contrastive learning, leveraging the label information of the input text records. It maximizes the agreements between causal text records belonging to the same class while minimizing agreements for causal text records from different classes. The formulation of the SuperCon loss is given as follows: Given a batch of $2N$ embedded records, z,

$$L = \sum_{i \epsilon I} L_i = \sum_{i \epsilon I} \frac{1}{|P_i|} \sum_{p \epsilon P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \epsilon A} \exp(z \cdot z_a/\tau)} \tag{4}$$

where i belongs to $I = 1, ..., N$ and represents the index of the anchor embedding z_i. The set of positive indices distinct from the anchor index i is denoted by $P_i \equiv p_i \in A(i) : y_p = y_i$, and $|P_i|$ is its cardinality. Here, y_p and y_i indicate the labels of the corresponding records. The scalar temperature parameter τ is used to scale the similarity measure.

In the loss calculation for a given batch, each record embedding z_i acts as an anchor embedding,

[1] Our code is available at https://github.com/rajathpatel23/causal-events.

[2] We use *roberta-base* model from Hugging Face model hub - https://huggingface.co/roberta-base

attempting to bring all record embeddings of the same class closer together in the embedding space while pushing away the record embeddings from different classes.

2.2 Fine-Tuning For Classification Task

In this phase, we leverage the pre-trained model from the first phase and adapt it specifically to the task of causal event classification.Fine-tuning is employed to optimize the model's parameters for this specific task, effectively utilizing its prior knowledge to enhance its ability to discern and categorize causal relationships within textual data.

To accommodate the classification task, we introduce a *Classification Head* atop the RoBERTa encoder

$$z = \text{RoBERTa}(x) \tag{5}$$

$$z = W_{ch}^{T} \cdot z + b_{ch} \tag{6}$$

where W_{ch} and b_{ch} are feed-forward layer specific weights and x is the input causal text. This is a simple single-layer feed-forward architecture. The primary purpose of this additional layer is to process the extracted embeddings and make predictions for the causal event classification. We employ the sigmoid activation function on the feed-forward output to derive the final probability

$$z_{out} = \sigma(z). \tag{7}$$

For training the model's parameters, we use binary cross-entropy loss, defined as follows:

$$J(\theta) = -\frac{1}{N}\sum_{i=0}^{N} \cdot y_i \cdot \log(z_{out})) + (1 - y_i) \cdot \log(1 - z_{out}) \tag{8}$$

where θ represents the parameters optimized during the fine-tuning phase, and y_i denotes the original labels for the causal input text records. The binary cross-entropy loss minimizes the difference between predicted and actual class assignments by comparing probabilities and true labels.

During fine-tuning, the encoder layer parameters are not frozen and fine-tuned end-to-end along with *Classification Head* parameters. This allows the model to specialize its learned representations for the causal event classification task without losing the valuable knowledge gained from pre-training.

Dataset	Causal	Non-Causal	Total
train	1624	1421	3075
dev	185	155	340
test	173	179	352

Table 1: Dataset distribution of Causal New Corpus

3 Experimentation Settings

3.1 Dataset

We utilize the Causal News Corpus, which is derived from the work of (Tan et al., 2022) for our experiments. This corpus is specifically prepared for the Shared Task on CASE 2023 Workshop on Event Causality Identification (Tan et al., 2023), focusing on Subtask 1 for causal event classification. This version contains more data than previous version of the dataset (Tan et al., 2022) while some previous annotations have been revised. The dataset comprises 869 news documents and 3767 English sentences that have been annotated with causal information. The corpus is partitioned into three sets: *train*, *dev*, and *test* splits to facilitate fair evaluation. A detailed distribution of the dataset can be found in Table 1.

3.2 Model Training

In the pre-training phase, we train the encoder architecture using the SuperCon Loss, with a batch size of 128. To guide the training process, we set the learning rate to 5e-5 and use a scalar temperature parameter, denoted as τ, which is set to 0.07. The pre-training runs for five epochs and involves both the *train* and *dev* splits from the causal news corpus dataset. To efficiently handle the data, we limit the maximum number of tokens for the encoder tokenizer to 256.

During the fine-tuning phase, we extend the pre-trained encoder architecture by adding a feed-forward network on top, known as the *Classification Head*. This additional network allows us to perform the specific task of causal event classification. We employ the binary cross-entropy loss (Eq. (8)) for training the model. Throughout fine-tuning, we solely use the *train* dataset and use the *dev* dataset to evaluate the model's performance. Finally, we submit the trained model's predictions on the *test* dataset to Codalab for evaluation on the hold-out test set. The parameters used for fine-tuning include - batch size of 16, the learning rate of 2e-5, and the number of training epochs set to 3, with an early stopping criterion.

User	Recall	Precision	F1-score	Accuracy	MCC
DeepBlueAI	0.8613 (5)	0.8324 (2)	0.8466 (1)	0.8466 (1)	0.6937 (1)
rpatel12	**0.8728 (4)**	**0.8162 (3)**	**0.8436 (2)**	**0.8409 (2)**	**0.6837 (2)**
timos	0.8786 (3)	0.8000 (4)	0.8375 (3)	0.8324 (3)	0.6683 (3)
csecudsg	0.8555 (6)	0.8000 (4)	0.8268 (4)	0.8239 (4)	0.6495 (4)
elhammohammadi	0.8960 (1)	0.7635 (6)	0.8245 (5)	0.8125 (5)	0.6352 (5)
tanfiona	0.8902 (2)	0.7586 (7)	0.8191 (6)	0.8068 (6)	0.6237 (7)
sgopala4	0.8613 (5)	0.7801 (5)	0.8187 (7)	0.8125 (5)	0.6288 (6)
nitanshjain	0.8728 (4)	0.6537 (8)	0.7475 (8)	0.7102 (8)	0.4483 (9)
kunwarv4	0.5260 (7)	0.8585 (1)	0.6523 (9)	0.7244 (7)	0.4819 (8)
pakapro	0.4740 (8)	0.4409 (9)	0.4568 (10)	0.4460 (9)	-0.1072 (10)

Table 2: The performance of the our model compared to all the other submission made to Codalab to CASE 2023 Shared Task 3 - Subtask 1 (Tan et al., 2023) on causal event classification

We manually select the hyper-parameters for the model during training. This approach ensures that the model's configuration aligns with the specific task requirements and contributes to its overall performance.

3.3 Evaluation Metrics

We employ various metrics, including Precision, Recall, F1-scores, Accuracy, and Matthew's correlation coefficient (MCC) to assess the performance of our binary classification model. Among these metrics, our model is optimized for the F1-score, which provides a balanced evaluation of both precision and recall.

4 Results and Analysis

This section presents the outcomes of our model architecture in the context of the causal event classification task. We conducted the model training on an A10 GPU with 24GB RAM, utilizing the available computational resources effectively.

4.1 Performance on Classification Task

Our contrastive learning based architecture is tailored for binary classification, determining if a given input text record x exhibits a semantic causal relationship. We compare its performance against other submissions in the event causality shared task 1 (Tan et al., 2023), summarized in Table 2. The results reveal our model's highly competitive performance in the classification task. It secures the 2nd position in three key metrics - F1-Score, Accuracy, and MCC. Additionally, it ranks 3rd in Precision and 4th in Recall among all submissions. Compared to the baseline model presented by (Tan et al., 2023), a fine-tuned BERT model with hyper-parameter tuning, our model shows significant improvements. Specifically, it achieves a remarkable 6-point increase in precision, a 3-point boost in F1-Score, and a substantial 6-point improvement in MCC score. These results provide strong evidence supporting the effectiveness of applying SuperCon learning to this specific classification problem.

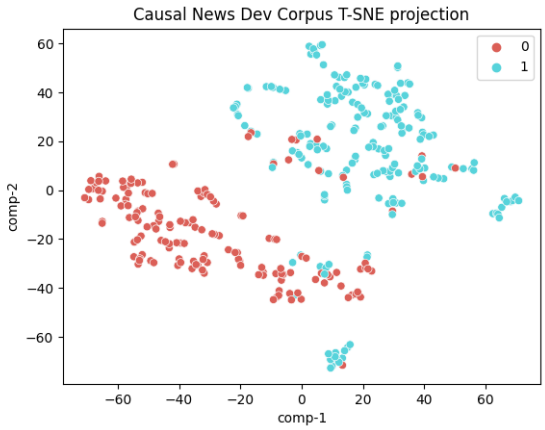

Figure 2: TSNE visualization of the representations from the pre-training phase

4.2 Analyzing Pre-trained Feature Spaces via t-SNE

To deepen our understanding of the impact of contrastive pre-training, we examine the feature representation generated from the *dev* dataset. The representation are visualized using the t-SNE technique (van der Maaten and Hinton, 2008). As depicted in Figure 2, the t-SNE plot reveals two clusters among the text records in the dataset. This clustering underscores the efficacy of our SuperCon-based pre-training approach. The visualization validates that the pre-training phase successfully imbues the model with meaningful representations, which, in turn, bolsters the model's performance in the causal event classification task. Interestingly, we observe

Model	Recall	Precision	F1-Score	Accuracy	MCC
BERT Baseline Model (Tan et al., 2023)	**0.887**	0.841	0.863	0.8471	0.6913
RoBERTa Non-Pre-trained Model	**0.9180**	0.8212	0.8181	0.8470	0.6941
Pre-trained OnlyModel	0.8756	0.7677	0.8673	0.7882	0.5755
Proposed SuperCon Model	0.8617	**0.8556**	**0.8972**	**0.8617**	**0.7210**

Table 3: Comparative study on the effectiveness of contrastive pretraining

some data point overlaps within the clusters, suggesting that these could be further refined through downstream tasks.

4.3 Effectiveness of Contrastive Pre-training

To comprehensively investigate the role of contrastive pre-training, we designed and executed experiments involving various model architectures. Specifically, we tested four different configurations:

BERT Baseline Model: This version uses the BERT architecture trained by (Tan et al., 2023) and serves as our foundational comparison point for the causal event classification task.

RoBERTa Non-Pre-trained Model: In this setup, we circumvent the pre-training phase altogether and train a RoBERTa encoder model with a classification head for the same combined number of epochs as our proposed model.

Pre-trained Only Model: In this scenario, the RoBERTa encoder model undergoes initial pre-training. During the fine-tuning stage, the feature-extracting layers are frozen, leaving only the classification head to be updated.

Proposed SuperCon Model: Our proposed architecture leverages the benefits of SuperCon Loss during the RoBERTa encoder model's pre-training phase, followed by a fine-tuning stage on the causal event classification task.

For a balanced comparative analysis, all model training was confined to the available *train* set, while evaluations were conducted on the *dev* dataset. The outcomes are summarized in Table 3.

The data reveal that our Proposed SuperCon Model excels in four metrics: Precision, F1-Score, Accuracy, and MCC, outperforming the other configurations. We also see a drop in performance metrics on the Pre-trained Only Model configuration, underscoring the necessity of fine-tuning subsequent to pre-training for achieving optimal results. Further the RoBERTa Non-Pretrained Model shows high recall but with lower F1-Score, Precision scores over our proposed model architecture.

5 Conclusion and Future Work

In this study, we have delved into the application of SuperCon learning for the task of causal event classification. By harnessing the power of SuperCon, our model achieved competitive performance, securing the 2nd position in key evaluation metrics such as F1-Score, Accuracy, and Matthew's correlation coefficient (MCC). These competitive results provide strong evidence for the efficacy of our approach in comprehending intricate causal relationships within textual data. Additionally, our comparative analysis highlights the model's learning strength and the benefits of this learning approach.

In the future we could explore the use of a large dataset from a distinct domain during the pre-training phase. This would enable us to gauge the inductive capacity of our learning paradigm on the causal news corpus domain dataset. Such investigations hold the potential for promising implications in the realms of low-resource, few-shot, and domain-specific causality event understanding.

Acknowledgements I thank members of Machine Learning Department at Interos Inc, including Britt Torrance, Hunter Powers, Rabia Turan for the support, compute resources, and their valuable comments and suggestions. I thank Saurabh Shringarpure for his insightful feedback. I also appreciate and acknowledge reviewers for their valuable comments and suggestions.

References

Alexander Brinkmann, Roee Shraga, and Christian Bizer. 2023. Sc-block: Supervised contrastive blocking within entity resolution pipelines. *arXiv*, abs/2303.03132.

J. Chen, Richong Zhang, Yongyi Mao, and Jie Xue. 2022. Contrastnet: A contrastive learning framework for few-shot text classification. *arXiv*, abs/2305.09269.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv*, abs/2002.05709.

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

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. *arXiv*, abs/1907.11692.

Zhiping Luo, W. Xu, Weiqing Liu, J. Bian, Jian Yin, and Tie-Yan Liu. 2021. Kge-cl: Contrastive learning of tensor decomposition based knowledge graph embeddings. In *International Conference on Computational Linguistics*.

Ralph Peeters and Christian Bizer. 2022. Supervised contrastive learning for product matching. In *Companion Proceedings of the Web Conference 2022*. ACM.

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Onur Uca, Thapa Surendrabikram, Farhana Ferdousi Liza, and Nelleke Oostdijk. 2023. Event causality identification with causal news corpus - shared task 3, CASE 2023. 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. 2022. 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.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. *arXiv*, abs/2204.05164.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172, Online. Association for Computational Linguistics.