ARGUABLY @ Causal News Corpus 2022: Contextually Augmented Language Models for Event Causality Identification

Guneet Singh Kohli
Thapar University, Patiala, India
guneetsk99@gmail.com

Prabsimran Kaur
Thapar University, Patiala, India
pkaur_be18@thapar.edu

Jatin Bedi
Thapar University, Patiala, India
jatin.bedi@thapar.edu

Abstract

Causal (a cause-effect relationship between two arguments) has become integral to various NLP domains such as question answering, summarization, and event prediction. To understand causality in detail, Event Causality Identification with Causal News Corpus (CASE-2022) has organized shared tasks. This paper defines our participation in Subtask 1, which focuses on classifying event causality. We used sentence level augmentation based on contextualized word embeddings of distillBERT to construct new data. This data was then trained using two approaches. The first technique used the DeBERTa language model, and the second used the RoBERTa language model in combination with cross attention. We obtained the second-best F1 score (0.8610) in the competition with Contextually Augmented DeBERTa model.

1 Introduction

Causality is a cause-effect relationship between two arguments, events, processes, states, or objects in which the occurrence of one (cause) is partly responsible for the occurrence of the other (effect) (Barik et al., 2016). A few instances of this cause-effect relationship are illustrated in Figure 1, which were extracted from the Causal News Corpus (CNC) (Tan et al., 2022a). The first instance comprises a causal relation between the phrase "allegedly being involved in the blast" (cause) and "Two more youths were arrested later," indicating that the youths were arrested because they were accused of being involved in the bomb blast. The word "for" indicates that this relationship is causal. Similarly, other words can be used for indication, as seen in the case of the second instance where "over" is the signal word. There are also cases where the causal relation is explicit and does not have a word to signal the causality, as can be seen in the third instance. For the sentences that do not have causality, they are either missing the effect or the cause argument missing (as illustrated by the fifth instance), or both.

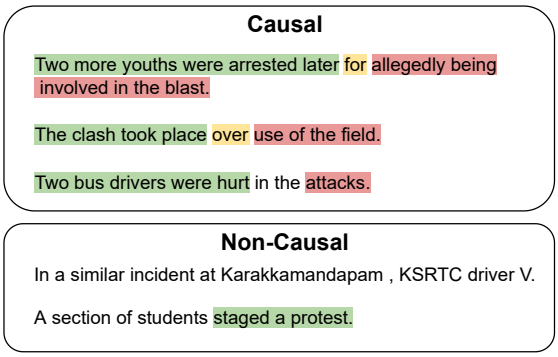

Figure 1: The cause, signal of causality, and effect are highlighted using the red, yellow, and green colors respectively. Any sentence that comprises of only cause or only effect is not considered as causal.

Causality is often used in Natural Language Processing (NLP) tasks that address Natural language inference and understanding (Jo et al., 2021; Dunietz et al., 2020; Feder et al., 2021). The information retrieved from the detection of causal relations can be used for various NLP tasks like Causal Question Answering and Generation applications (Dalal et al., 2021; Hassanzadeh et al., 2019; Stasaski et al., 2021), and Event prediction (Radinsky et al., 2012). However, identifying and extracting a causal relationship is challenging as it requires significant semantic knowledge.

This paper describes our participation in the Event Causality Identification with Causal News Corpus (CASE-2022), the third shared task of the CASE 2022 (Tan et al., 2022b). Under this task, there are two subtasks, and this paper describes an approach for subtask 1. We have used the following methods to classify causal events:

- We used sentence level augmentation based on contextualized word embeddings of distillBERT to construct new data.
- The training of this data is done using two approaches. The first technique used the De-

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 143 - 148

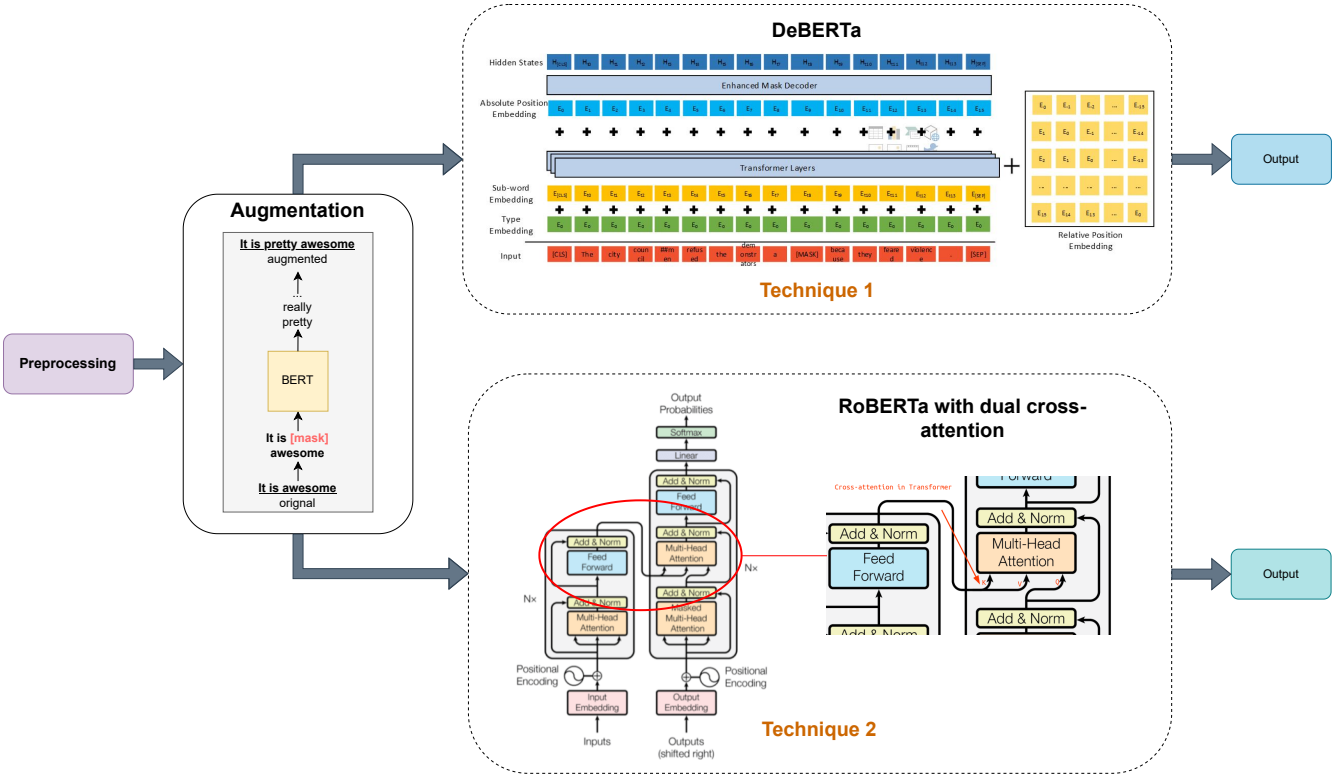

Figure 2: Architecture of the proposed pipeline. The initial part of the pipeline is same for both the techniques. Note that we illustrate the Encoder portion of RoBERTa with dual cross attention. The other components of RoBERTa architecture were not refactored for any changes

BERTa language model, and the second used the RoBERTa language model in combination with cross-attention.

The aim of the task and the details of data is explained in Section 2. Section 3 gives a detailed overview of the method used for the binary classification. The results obtained, it's analysis and the experimental setup is described in Section 4.

2 Task & Data Description

Event Causality Identification Shared Task aims at tackling inference and understanding by organizing two subtasks: a) Causal Event Classification and b) Cause-Effect-Signal Span Detection. Our team participated in the first subtask, which required the participants to classify the given text into "0" (non-causal) and "1"(causal). The dataset provided in the task was the Causal News Corpus (CNC) deals with event causality in the news. The CNC dataset builds upon the following datasets: Automated Extraction of Socio-political Events from News (AESPEN) in 2020 (Hürriyetoğlu et al., 2020b,a) and Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) in 2021 (Hürriyetoğlu et al., 2021a,b). The data in

Label	Train	Dev
0	1322	145
1	1603	178

Table 1: Data distribution for the CNC database.

CNC is based on random samples that have been collected from a total of 869 news documents. The corpus comprises 3,559 data samples, out of which 2925 data points were provided for training, 323 data points were provided for the development set, and the remaining 311 samples were used as the test set.

3 Methodology

This section gives an exhaustive overview of the proposed pipeline. Section 3.1 provides the details of the preprocessing performed on the given dataset. Section 3.2 describes the augmentation technique used on the data. Section 3.3 discusses the transformer models and techniques used to train the data.

Orignal	Augmented	Label
The protests spread to 15 other towns and resulted in two deaths and the destruction of property .	the protests **had** spread **quickly** to 15 other towns and **resulted ultimately in** two **premature** deaths **locally** and **essentially** the destruction of property.	**1**
The protests spread to 15 other towns and resulted in two deaths and the destruction of property .	the protests spread **out** to **over** 15 other **other** towns **offshore and territory** resulted in two **workers** deaths and the destruction of property	**1**
The demonstrations pose a real problem , not just for the British but for others too.	the demonstrations pose **considered a serious** real **negativ**e problem, **affecting** not just **resentment** for the british but for others all too.	**0**
The demonstrations pose a real problem , not just for the British but for others too.	the demonstrations **allegedly** pose **potentially** a real **world** problem, not just **perhaps** for **interested** the british but for **important** others too.	**0**

Table 2: Illustration of the contextual augmentation performed by our proposed methodology. The words that are added or changed in the original sentence have been highlighted.

Model	Recall	Precision	F1	Accuracy	MCC
Top Performing	0.8864	0.8387	0.8619	0.8392	0.6714
Proposed Model	**0.9148**	**0.8131**	**0.8610**	**0.8328**	**0.6602**
Average Score	0.8686	0.7838	0.8233	0.7870	0.5619

Table 3: Comparison of the proposed model with the top performing model and the average results of all the models on the leaderboard. The proposed model refers to our best-performing model DeBERTa trained on augmented data.

3.1 Data Pre-Processing

The quality of data highly impacts the performance of any machine learning or deep learning model. However, the raw data present is unstructured. It comprises noise, punctuations, special symbols, and unusual texts that might affect the feature selection process of the model, causing it to underperform. Thus a basic preprocessing involving the tokenization of the sentences into words, conversion of the words into lowercase, and removal of stopwords (the, an, a) and punctuations was done using the NLTK library (Loper and Bird, 2002).

3.2 Augmentation

Models like BERT and RoBERTa comprise millions of parameters that require a considerable amount of data to generalize and obtain meaningful results. However, the dataset provided to the participants has only 2925 data points, which is insufficient to train these heavy models. Thus, data augmentation, a technique of applying transformation on the original labeled data to construct new data , was used on the training data to reduce overfitting. NLPaug tool[1], a well-known library that can perform three types of augmentations: Character level, Word Level, and Sentence Level was used for augmentation in our pipeline. For this task, we used sentence-level augmentation based on contextualized word embeddings of distillBERT. NLPaug also allows you to perform various actions like 'Insertion' and 'Substitution' operations. Our technique utilizes the Insertion operation, which randomly picks a position in the sentence and inserts in that position a word that best fits the local context. It was ensured that the causality of the dataset was not changed during the augmentation process as can be seen in Table 2. Contextualized word embeddings provide these words chosen for insertion.

3.3 Modeling

A transformer-based approach is used to perform Event Causality Identification. The training was done using DeBERTa (He et al., 2020) and dual Cross attention RoBERTa (Liu et al., 2019). A detailed explanation of their architecture is given in this Section. The architecture of the pipeline is illustrated in Figure 2.

3.3.1 DeBERTa

Decoding-enhanced BERT with Disentangled Attention (DeBERTa) is an enhanced version of the

[1]urlhttps://github.com/makcedward/nlpaug.

BERT and RoBERTa. It differs from BERT in two aspects. The first is the disentangled self-attention mechanism, which involves using two vectors to encode the content and position rather than a single vector to address these embeddings. This helps the model naturally encode the word position information, which conventional transformers lack. The second is Enhanced Mask Decoder (EMD), a technique that performs masked token prediction in model pre-training using absolute positions in the decoding layer, unlike BERT, which uses relative position. This helps DeBERTa obtain better accuracy since the words' syntactic roles depend highly on their absolute positions in a sentence.

3.3.2 Dual Cross attention RoBERTa

Robustly Optimized BERT-Pretraining Approach (RoBERTa) is an extension of BERT. Similar to BERT, data is passed through RoBERTa in the form of sequences. However, before passing these sequences, they are tokenized into words, the sequences are masked, a [CLS] token is added to the beginning of the first sentence, and a [SEP] is added after each sequence to indicate the end. Three embeddings, namely, token, sentence, and positional, are attached to each token. Once the encoding is done, these sentences are passed through the transformer.

RoBERTa differs from BERT in the aspect of token masking. BERT used a static masking technique while pretraining, in which each sequence was masked in 10 different patterns. The training data was further trained for 40 epochs indicating that each sequence was trained for the same masking pattern four times. Unlike BERT, RoBERTa was trained using a dynamic masking technique where a new masking pattern is generated every time a sequence is fed into the model. This helps create a more generalised model.

In the proposed pipeline, we used two layers of cross-attention while training RoBERTa to enhance the overall performance. In contrast to the conventionally used self-attention technique, which takes a single embedding sequence as input, the cross-attention combines two different asymmetrical sequences of identical dimensions. One of the sequences serves as a query input, while the other as a key and value input.

4 Results and Discussion

4.1 Comparative Analysis

In this section we present a detailed comparison of our best submission with other submissions present on the leaderboard.The comparitive study can be observed in Table 3. Our system ranked 2nd overall with F1 Score of 0.8610. The following results were obtained with **DeBERTa trained on Augmented data with Token length of 450**. The contextualized word embedding augmentation with distillBERT helped DeBERTa be more robust and handle the test data well. The best performing system of the task had F1 score of 0.0009 greater than our submission. Our system reports the highest Recall of 0.9148 across the leaderboard. The high recall is a direct indicator of high quality of augmented data we had produced for the task. In comparison to the average scores calculated from the leaderboard our system had 4.5% higher F1 score, 5.3% higher recall and 5.819% higher accuracy.

4.2 Experimental Setup

We trained the language models on Tesla-T4 16 GB GPU. For training, we kept the batch size as four and configured the AdamW optimizer with the learning rate of 1e-05. We fine-tuned the language models with a token length of 450 and trained the data up to 3 epochs.

4.3 Analysis of Experiments

This section discusses the results and performance of our models, DeBERTa and Dual Cross Attention RoBERTa, as illustrated in Table 4. The core idea was the introduction of contextual augmentation using fine-tuned distillBert. The use of contextual embedding helped maintain the causality of the sentence that was necessary for the scope of the task. The increase in the data significantly impacted the performance of the proposed approaches. DeBERTa fine-tuned on augmented data yielded an F1 score of 0.8610 [our best performing system], an improvement of 3.5% from the unaugmented data version. For Dual Cross Attention RoBERTa, using augmented data brought about a gain of 2.6%.

DeBERTa uses disentangled attention which computes the attention weight of a word pair as a sum of four attention scores using disentangled matrices on their contents and positions as content-to-content, content-to-position, position-to-content, and position-to-position.

Model	Recall	Precision	F1	Accuracy	MCC
RoBERTa [Naive] unaugmented,Token length:450	**0.9261**	0.7477	0.8274	0.7813	0.5615
RoBERTa [Dual Cross Attn] unaugmented,Token length:450	0.8806	0.7868	0.8311	0.7974	0.5858
RoBERTa [Dual Cross Attn] augmented,Token length:450	0.8977	0.8061	0.8494	0.8199	0.6327
DeBERTa unaugmented,Token length:450	0.8863	0.7839	0.8320	0.7974	0.5862
DeBERTa augmentated,Token length:450	0.9148	**0.8131**	**0.8610**	**0.8328**	**0.6602**

Table 4: Results of the models experimented on. The Best Performing System has been highlighted.

Text	Gold	Predicted
Rath interacted with the affected farmers who were yet to get compensation despite repeated agitation over the issue .	0	1
Another ‘ TP ’ issue may also leave a blot on the CPM , as public opinion is heavily pitted against the assault made upon former diplomat T P Srinivasan by SFI activists .	0	1
Police said fighting broke out in Charbatan area in Murshidabad constituency even as the results were being declared .	0	0
Some protesters attempted to fight back with fire extinguishers.	0	0
The one-day fast attracted a " motley crowd " according to Sumitra M. Gautama, a teacher with the Krishnamurthi Foundation of India (KFI)	1	0
Both sides were raining bombs on each other and Mondal was hit by one of the bombs , " Murshidabad district magistrate Pervez Ahmed Siddiqui said .	1	0
SI Gopal Mondal , who was part of the police team that rushed to the spot , was killed by a crude bomb explosion .	1	1
The workers had embarked on a wildcat strike demanding better working conditions .	1	1

Table 5: Behavioural Analysis of our best performing model (DeBERTa with augmentation) on the validation set.

The position-to-content term is impactful since the attention weight of a word pair depends not only on their contents but on their relative positions, which is calculated by the content-to-position and position-to-content terms. The causality of a sentence is highly sensitive to the positioning of words in the sentence, and thus DeBERTa uses the position-to-content weights to capture the underlying semantics of the causality.

We used dual cross attention in RoBERTa by generating two embedding representation of an instance and calculating the attention weights for generating the attention filters. The improvements in the results can be observed in Table 4.

4.4 Quantitative analysis

This section discusses the quantitative analysis of the labels predicted by our best model on the validation set. Table 5 illustrates a few instances from the validation dataset along with the original and predicted labels. The first two instances demonstrate the cases where the model failed to understand the semantic meaning of words like "affected," "issue," and "against" and interpreted the immediate sense rather than trying to understand what the sentence as a whole means.

The fifth and sixth instance demonstrates the model's inability to distinguish the cause and effect portions of the sentence. "The one-day fast attracted a motley crowd " was considered the cause and thus could not find any effect, thus predicting this sentence to be non-causal. Similarly, the model did not identify "was hit by one of the bombs" as the effect for the sixth instance. Instances three, four, seven, and eight, demonstrate the cases where the model successfully understood the semantics and identified the cause-effect relations.

5 Conclusion

In this paper we propose Contextual Embedding Augmented DeBERTa and Dual Cross Attention RoBERTa to identify event causality. Our approach yielded an F1 score of 0.8610 which was the second best system throughout the shared task. We study the behaviour of both the models in augmented and unaugmented settings to derive proper understanding about the impact of our complete pipeline. In future, experimenting with other language models and extensive hyperparameter tuning through Neural Architectural Search will be an ideal path to follow. The augmentation was successful in maintaining its causality nature. This acts as a fine way of up sampling low resource tasks which lack adequate data.

References

Biswanath Barik, Erwin Marsi, and Pinar Øzturk. 2016. Event causality extraction from natural science literature.

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

Jesse Dunietz, Gregory Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and David Ferrucci. 2020. To test machine comprehension, start by defining comprehension. *arXiv preprint arXiv:2005.01525*.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2019. Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. In *IJCAI*, pages 5003–5009.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Farhana Ferdousi Liza, Ritesh Kumar, and Shyam Ratan. 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.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, and Erdem Yörük. 2021b. Challenges and applications of automated extraction of socio-political events from text (case 2021): Workshop and shared task report. *arXiv preprint arXiv:2108.07865*.

Ali Hürriyetoğlu, Erdem Yörük, Vanni Zavarella, and Hristo Tanev. 2020a. Proceedings of the workshop on automated extraction of socio-political events from news 2020. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*.

Ali Hürriyetoğlu, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020b. Automated extraction of socio-political events from news (aespen): Workshop and shared task report. *arXiv preprint arXiv:2005.06070*.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

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 preprint arXiv:1907.11692*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918.

Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A Hearst. 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.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022a. The causal news corpus: Annotating causal relations in event sentences from news. *arXiv preprint arXiv:2204.11714*.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Hansi Hettiarachchi, Tadashi Nomoto, Onur Uca, and Farhana Ferdousi Liza. 2022b. 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 2022)*, Online. Association for Computational Linguistics.