CSECU-DSG @ Causal News Corpus 2022: Fusion of RoBERTa Transformers Variants for Causal Event Classification

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{aziz.abdul.cu, akram.hossain.cse.cu}@gmail.com,
and nowshed@cu.ac.bd

Abstract

Identifying cause-effect relationships in sentences is one of the formidable tasks to tackle the challenges of inference and understanding of natural language. However, the diversity of word semantics and sentence structure makes it challenging to determine the causal relationship effectively. To address these challenges, CASE-2022 shared task 3 introduced a task focusing on event causality identification with causal news corpus. This paper presents our participation in this task, especially in subtask 1 which is the causal event classification task. To tackle the task challenge, we propose a unified neural model through exploiting two fine-tuned transformer models including RoBERTa and Twitter-RoBERTa. We perform score fusion through combining the prediction scores of each component model using weighted arithmetic mean to generate the probability score for class label identification. The experimental results showed that our proposed method achieved the top performance (ranked 1st) among the participants' systems.

1 Introduction

Causality is a fundamental cognitive concept that frequently emerges in various natural language processing (NLP) works. It mostly focuses on the challenges of inference and understanding of the natural language. In general, a causal relation is a semantic relationship between two arguments known as cause and effect, where the occurrence of one (cause argument) incurs the occurrence of the other (effect argument). Such causal relation plays an important role in various contemporary NLP tasks including document-summarization, event prediction from text, scene and story generation, question-answering (Q/A), product recommendation based on user comments, and other textual entailments (Yu et al., 2022; Yang et al., 2022).

**The first two authors have equal contributions.

To address the challenges of event causality identification in texts, Tan et al. (Tan et al., 2022a) introduced a shared task at the CASE-2022 workshop. The task is composed of two subtasks including a causal event classification task (subtask1) and a cause-effect-signal span detection task (subtask 2). However, we only participated in the causal event classification task (subtask1), where given a text a system needs to determine whether it contains a cause-event meaning or not. To demonstrate a clear view of the task definition, we articulate a few examples from subtask 1 in Table 1.

Sentence	Label
The farmworkers ' strike resumed on Tuesday when their demands were not met	1
He said he was about 100 metres away when he witnessed the attack .	0

Table 1: Example of subtask 1. Here, label 1 means Causal and 0 means Non-Causal.

Prior work on event causality identification has mostly employed semi-supervised methods (Rink et al., 2010; Mirza, 2014; Aziz et al., 2020) based on features (e.g. psycho-linguistic, syntactic, semantic, etc.) or supervised methods (Gordeev et al., 2020; Ionescu et al., 2020) based on transformers model (e.g. BERT, RoBERTa, DistilBERT, etc.). However, transformer-based methods obtained more competitive results (Mariko et al., 2020), although those methods have some limitations in the fusion technique. In order to overcome this limitation, we proposed a RoBERTa-based unified method where we utilise the weighted average fusion technique.

We organize the rest of the paper as follows: Section 2 describes our proposed system in the CASE-2022 causal event classification task whereas, in

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 138 - 142

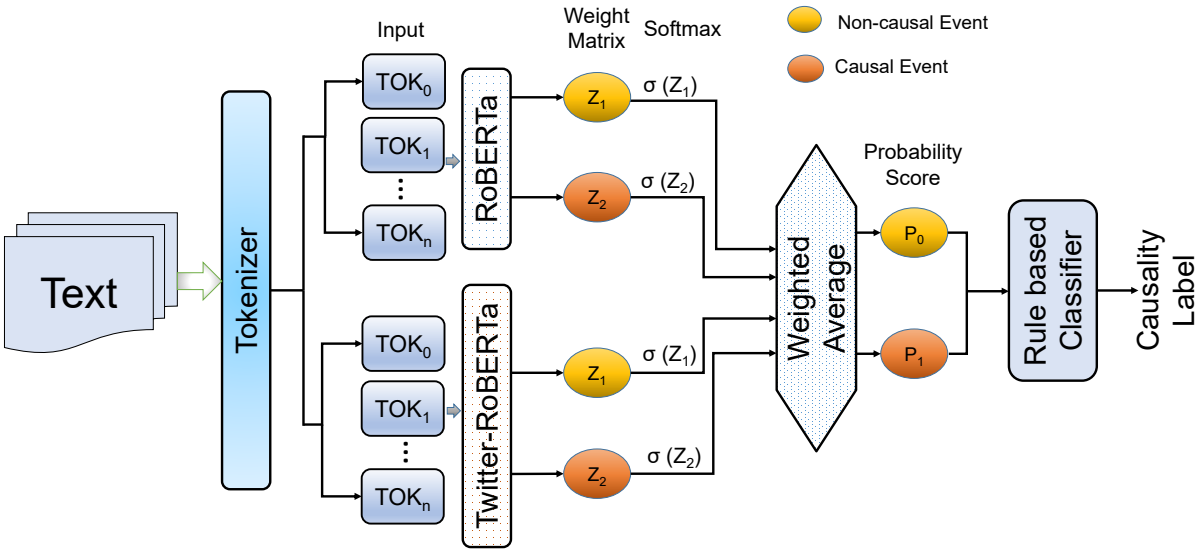

Figure 1: Our proposed model for the causal event classification task.

Section 3, we present our system design with parameter settings and conduct the results and performance analysis. Finally, we conclude with some future directions in Section 4.

2 Proposed Framework

In this section, we describe our proposed approach for the event causality identification task. Our goal is to exploit the inherent semantics of the sentence to identify whether the event sentence contains any cause-effect meaning. The overview of our proposed framework is depicted in Figure 1.

Given an input text, we employ two transformer models including RoBERTa (Liu et al., 2019) and one of its variants Twitter_RoBERTa (Barbieri et al., 2020) to extract the diverse contextual features. Such feature representation better captures the inherent semantics of the text. Later, a linear feed-forward layer is utilized in each model to estimate the probability score of each class. Finally, for the effective fusion of the scores, we take the weighted arithmetic mean of the prediction scores of these models. A class that contains the highest probability scores is considered as the final label.

2.1 Transformer Models

RoBERTa (Liu et al., 2019) stands for robustly optimized BERT pre-training approach. RoBERTa has the same architecture as BERT, but it eliminates the next sentence prediction (NSP) objective used in BERT during pre-training. Besides, it trained on longer sequences with much larger mini-batches and learning rates. Instead of using static masking like BERT, RoBERTa utilizes dynamic masking that is employed every time a text sequence is fed to the model. Therefore, the model encodes the several versions of the same sentence with masks on different positions. It helps the model to capture the inherent semantics of the text.

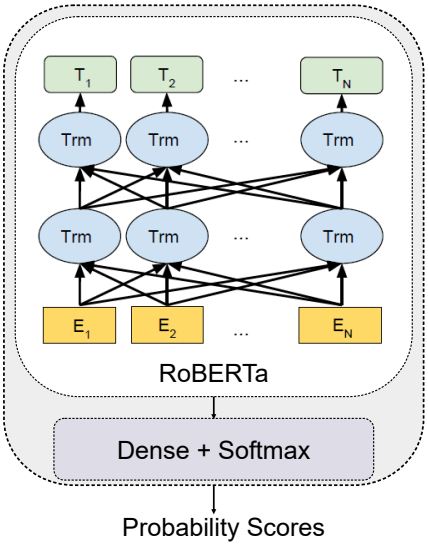

Figure 2: RoBERTa model.

We also employ the Twitter_RoBERTa (Barbieri et al., 2020), a RoBERTa-base model trained on 58M tweets, described and evaluated in the TweetEval benchmark. In our proposed framework, we use RoBERTa along with its Twitter variants to capture the diverse semantic features effectively. Here, we use the HuggingFace's implementation of the *roberta-base* model (Wolf et al., 2019). It is composed of 12-layers (i.e. transformer block), the

dimension of hidden size is 768, the number of the self-attention head is 12, and contains 125M parameters. In Figure 2, we demonstrate an overview of the setup of RoBERTa transformer model to obtain the prediction score of each text.

2.2 Fusion of Transformer Models

In the NLP domain, it is usually a common practice to fuse multiple models to enhance the performance of individual models or tackle the limitations of models. In our proposed framework, we also employ a fusion strategy to combine the effectiveness of RoBERTa and Twitter_RoBERTa transformer models. We estimate a unified probability score for each class through fusing the prediction scores generated from each model. For the score fusion, we employ the weighted arithmetic mean of these two scores. Finally, based on the highest probability score, we determine the final label for a given text. The estimation is computed as follows:

$$f(x_i, y_i) = \begin{cases} 0, & \text{if } W_0 > W_1 \\ 1, & \text{otherwise} \end{cases}$$

$$W_i = \frac{x_i * R + y_i * T}{R + T} \quad (1)$$

x_i and y_i correspond to the RoBERTa and Twitter-RoBERTa probability score, where R and T represent their weight respectively. W_i (i.e. i = {0, 1}) is the unified probability score for each class.

3 Experiment and Evaluation

3.1 Dataset Description

The organizers used the Causal News Corpus (CNC) (Tan et al., 2022b), a benchmark dataset published in LREC-2022 to evaluate the performance of the participants' systems at the CASE-2022 event causality shared task (Subtask 1). The dataset statistics are summarized in Table 2.

Category	Causal	Non-Causal	Total
Train	1603	1322	2925
Dev	178	145	323
Test	176	135	311
Total	1957	1602	3559

Table 2: The statistics of causal news corpus used in event causality shared task in CASE-2022.

The dataset comprises of 3559 event sentences where 2925, 323, and 311 samples are used for the train, dev, and test phases. Each sentence is annotated with binary labels (Causal: 1 and Non-Causal: 0) which indicates whether there is a causal relationship available in a sentence or not.

3.2 Experimental Settings

We now describe the details of our experimental settings and the hyper-parameter settings with fine-tuning strategy that we have employed to design our proposed CSECU-DSG system for the CASE-2022 event causality identification shared task.

Parameter	Optimal Value
Learning rate	3e-5
Max-len	128
Epoch	5
Batch size	16
Manual seed	4

Table 3: Model settings for CASE-2022 event causality identification shared task (subtask 1).

In our CSECU-DSG system, we utilize two state-of-the-art Huggingface transformer models with fine-tuning, including RoBERTa and Twitter-RoBERTa. We use simpletransformers API (Rajapakse, 2019) to implement our system. We use the train and development data during the model training phase. We used the CUDA-enabled GPU and set the manual seed = 4 to generate reproducible results. We obtained the optimal parameter settings of our proposed model based on the performance of the development set which articulated in Table 3 and we used the default settings for the other parameters.

To generate the unified prediction, we fuse the probability score of RoBERTa and Twitter-RoBERTa based classification model as described in Section 2.2. To select the optimal weight as defined in Equation 1, we swept the parameter value of R and T between $\{0.1, ..., 0.9\}$ and conduct some experiments on training data. Based on the experimental results, we choose the weight $R = 0.6$ for RoBERTa and weight $T = 0.4$ for Twitter-RoBERTa model.

3.3 Evaluation Measures

To evaluate the participants' system at the CASE-2022 event causality identification shared task (sub-

https://huggingface.co/roberta-base
https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Team (Rank)	Recall	Precision	F1-score	Accuracy	MCC
CSECU-DSG (1st)	0.8864	0.8387	0.8619	0.8392	0.6714
Participants system performance on subtask 1					
Arguably (2nd)	0.9148	0.8131	0.8610	0.8328	0.6602
hiranmai (3rd)	0.8864	0.8211	0.8525	0.8264	0.6451
NLP4ITF (4th)	0.8807	0.8245	0.8516	0.8264	0.6449
IDIAPers (6th)	0.8750	0.8280	0.8508	0.8264	0.6449
LXPER AI Research (9th)	0.8636	0.8261	0.8444	0.8199	0.6318
Innovators (15th)	0.7898	0.7202	0.7534	0.7074	0.3981
Baseline	0.8466	0.7801	0.8120	0.7781	0.5452

Table 4: Comparative results with other selected participants (Subtask 1).

Method	Recall	Precision	F1-score	Accuracy	MCC
CSECU-DSG	0.8864	0.8387	0.8619	0.8392	0.6714
Performance of individual model					
RoBERTa	0.8807	0.8245	0.8516	0.8264	0.6449
Twitter-RoBERTa	0.8409	0.8087	0.8245	0.7974	0.5858

Table 5: Performance analysis of individual model used in our proposed CSECU-DSG system (Subtask 1).

task 1) (Tan et al., 2022a), the organizers employed standard evaluation metrics including recall, precision, F1-score, accuracy, and Matthews correlation coefficient (MCC) (Matthews, 1975). However, the F1 score is considered as the primary evaluation metric for subtask 1 and systems performances were ranked based on this score.

3.4 Results and Analysis

In this section, we analyze the performance of our proposed CSECU-DSG system in the CASE-2022 event causality identification shared task (subtask 1). The comparative results of our proposed CSECU-DSG system along with other top-performing systems (Tan et al., 2022a) in subtask 1 are presented in Table 4. Following the benchmark of CASE-2022 event causality identification subtask 1, participants' systems are ranked based on the primary evaluation metric F1 score.

At first, we presented the performance of our proposed CSECU-DSG system. We also presented the performance of top-ranked participating systems and the baseline used in subtask 1. Here, we see that our proposed method obtained the highest score in terms of the primary evaluation metric F1 score compared to the other top-performing systems. This deduces the superiority and effectiveness of our proposed system for the event causality identification task.

In our proposed CSECU-DSG system, we perform the effective fusion of two state-of-the-art RoBERTa transformer models. However, to validate the performance of our used fusion strategy, we conduct individual experiments using each transformer models to estimate the effect of each model used in our proposed system. The summarized experimental results regarding this are presented in Table 5.

From the results, it can be observed that RoBERTa based model performed better compared to the Twitter-RoBERTa model when considering individual model performances. However, combining two models prediction scores by using weighted arithmetic mean improved the performance. It shows that the fusion strategy improves the $\sim$1% performance compared to the RoBERTa model and improves the $\sim$4% performance compared to the Twitter-RoBERTa model in terms of the primary evaluation measure F1 score. This validates the importance of our fusion strategy.

4 Conclusion and Future Directions

In this paper, we present an approach to identify the cause-effect relation in texts by exploiting RoBERTa variants with an effective fusion strategy. Experimental results demonstrated the efficacy of our fusion strategy of the two SOTA transformers model which helped us to obtain the best result in subtask 1.

In the future, we intend to explore other indicators of textual causal relations for further improvement. Especially, a graph-based neural model may exploit complex dependency patterns of cause-effect relations from text more effectively.

References

Abdul Aziz, Afrin Sultana, Md Akram Hossain, Nabila Ayman, and Abu Nowshed Chy. 2020. Feature fusion with hand-crafted and transfer learning embeddings for cause-effect relation extraction. In *FIRE (Working Notes)*, pages 756–764.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Denis Gordeev, Adis Davletov, Alexey Rey, and Nikolay Arefyev. 2020. Liori at the fincausal 2020 shared task. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 45–49.

Marius Ionescu, Andrei-Marius Avram, George-Andrei Dima, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Upb at fincausal-2020, tasks 1 & 2: Causality analysis in financial documents using pretrained language models. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 55–59.

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*.

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

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

Paramita Mirza. 2014. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 10–17.

T. C. Rajapakse. 2019. Simple Transformers. `https://github.com/ThilinaRajapakse/simpletransformers`.

Bryan Rink, Cosmin Adrian Bejan, and Sanda Harabagiu. 2010. Learning textual graph patterns to detect causal event relations. In *Twenty-Third International FLAIRS Conference*.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Hansi Hettiarachchi, Tadashi Nomoto, Onur Uca, and Farhana Ferdousi Liza. 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 2022)*, Online. 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 Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, pages 1–26.

Xiaoxiao Yu, Xinzhi Wang, Xiangfeng Luo, and Jianqi Gao. 2022. Multi-scale event causality extraction via simultaneous knowledge-attention and convolutional neural network. *Expert Systems*, 39(5):e12952.