CSECU-DSG @ Causal News Corpus 2023: Leveraging RoBERTa and DeBERTa Transformer Model with Contrastive Learning for Causal Event Classification

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

Abstract

Cause-effect relationships play a crucial role in human cognition, and distilling cause-effect relations from text helps in ameliorating causal networks for predictive tasks including natural language-based financial forecasting, text summarization, and question-answering. However, the lack of syntactic clues, the ambivalent semantic meaning of words, and complex sentence structures make it one of the challenging tasks in NLP. To address these challenges, CASE-2023 introduced a shared task 3 with two subtasks focusing on event causality identification with causal news corpus. In this paper, we demonstrate our participant systems for this task. We leverage two transformers models including DeBERTa and Twitter-RoBERTa along with the weighted average fusion technique to tackle the challenges of subtask 1 where we need to identify whether a text belongs to either causal or not. For subtask 2 where we need to identify the cause, effect, and signal tokens from the text, we proposed a unified neural network of DeBERTa and DistilRoBERTa transformer variants with contrastive learning techniques. The experimental results showed that our proposed method achieved competitive performance among the participants' systems and achieved 4th and 3rd rank in subtasks 1 and 2 respectively.

1 Introduction

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). Causal relation extraction from text is also known as the study of causality extraction (CE) which gain attention in different domains including Biomedical, media, emergency management (Bui et al., 2010; Balashankar et al., 2019; Qiu et al., 2017), etc. Such causal relation plays an important role in various contemporary NLP tasks including question-answering (Q/A), product recommendation based on user comments, and other textual entailments (Yu et al., 2022; Yang et al., 2022). However, the implicit causal relationship between sentences, numerical connectives impact, and ambivalent semantic meaning of the text make CE one of the most challenging tasks in NLP.

Subtask 1	
Sentence	Label
He said he was about 100 metres away when he witnessed the attack.	0
It has organised a political convention to mobilise support to secular forces.	1
Subtask 2	
Text: In 2009, riots broke out in the capital, Urumqui, and in their wake, mass arrests were made and many Uyghurs were imprisoned. **Label:** O O B-C I-C I-C I-C I-C I-C O O B-S I-S I-S B-E I-E I-E I-E I-E I-E I-E I-E I-E	

Table 1: Example of sub-task 1 and subtask 2 where subtask 2 labels are converted into BIO format, C = Cause, E = Effect, and S = Signal.

To address these challenges of event causality identification in texts, Tan et al. (2023) introduced a shared task 3 at the CASE-2023 workshop. The task is composed of two subtasks including a causal event classification task (subtask 1) and a cause-effect-signal span detection task (subtask 2). In subtask 1, participants ask to build an automatic system to classify a given text whether it contains a causal event meaning or not. Subtask 2 introduce different challenges for participant it aims to identify the cause, effect, and signal spans of that given text. To demonstrate a clear view of the task definition, we articulate a few examples from Subtask 1 and Subtask 2 in Table 1.

Prior work on event causality identification has

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 108–112, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_015

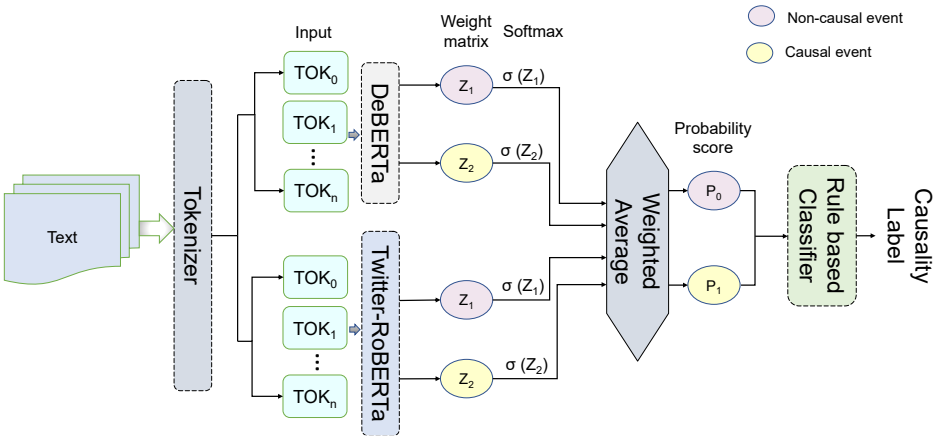

Figure 1: Our proposed model for subtask 1.

mostly employed semi-supervised methods (Mirza, 2014) based on features (e.g. psycho-linguistic, syntactic, etc.) or supervised methods (Gordeev et al., 2020; Ionescu et al., 2020) based on transformers model (e.g. BERT, RoBERTa, etc.). Though, transformer-based methods obtained more competitive results (Ionescu et al., 2020; Mariko et al., 2022), but those methods are either well performed for subtask 1 or subtask 2 problems but limited to well performed on both problems at the same time. In order to overcome this limitation, we proposed generalized architecture for both types of tasks. Where we fuse two different transformers models including DeBERTa and Twitter RoBERTa or DistilRoBERTa with different fusion techniques. We utilize the prediction level late fusion technique for subtask 1 whereas, for subtask 2 we use the feature level early fusion technique. Although these switching in place of transformers and fusion techniques help us to achieve competitive results in the competition. Moreover, we utilized unsupervised contrastive learning to address the spans section more precisely for subtask 2.

Accordingly, the remaining sections of the paper are organized as follows: Section 2 introduces our proposed system in CASE 2023 for automatically identifying causal events from given text, while Section 3 presents our system design, parameter settings, and primary evaluation measures. Additionally, in this section, we also discuss our results and performance analysis. Finally, we conclude with some future directions in Section 4.

2 Proposed Method

In this section, we describe our proposed approach for CASE 2023 task 3, subtask 1 and subtask 2.

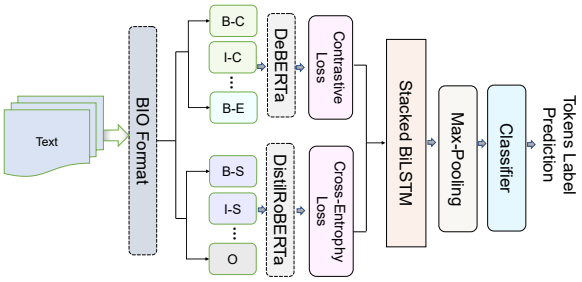

Figure 2: Our proposed model for subtask 2.

The overview of our proposed framework for subtask 1 is depicted in Figure 1. To extract the diverse contextual features from the text, we employ two transformer models including DeBERTa (He et al., 2021a) and one of RoBERTa variants Twitter_RoBERTa (Barbieri et al., 2020). 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 the final label.

On the other hand for subtask 2 we utilized two different transformer models DeBERTa and DistilRoBERTa independently to exploit cause-effect and signal span features respectively. Then we concatenate both transformers model features and feed to a stacked BiLSTM network to distill long-term relations among the tokens. Followed by the BiLSTM network we incorporate a max-pooling and classifier layer to predict tokens label. To improve system performance we calculate the contrastive loss for cause-effect token classification whereas we utilized cross-entropy loss for signal token classification since it may or may not contain in text.

However, Figure 2 illustrates our proposed method for subtask 2.

2.1 Transformer Model

DeBERTa[1] (He et al., 2021b) stands for decoding-enhanced BERT with disentangled attention. It improves the BERT and RoBERTa models using disentangled attention mechanism and enhanced mask decoder. We used the enhanced version of the DeBERTa model named DeBERTaV3 (He et al., 2021a). To improve the DeBERTa model, the DeBERTaV3 model used ELECTRA style pre-training where replacing mask language modeling (MLM) with the replaced token detection (RTD). It also used the gradient-disentangled embedding sharing (GDES) method to share the embeddings with the discriminator. These significantly improved the performance of the DeBERTa model in downstream tasks. Motivated by this, we employ Huggingfaces' (Wolf et al., 2019) implementation of *microsoft/deberta-v3-base* checkpoint to extract the feature representations of the sentence.

We also employ the Twitter_RoBERTa[2] (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 its to capture the diverse semantic features from short input text effectively. Moreover, in subtask 2 we used another transformer model DistilRoBERTa to extract word-based contextual representation to learn low-level features from the text. However, our experiments finds that DistilRoBERTa performed well in subtask 2 compare with Twitter RoBERTa. We utilize DistilRoBERTa base[3] (Sanh et al., 2019) model which is finetuned on conell-03 dataset.

2.2 BiLSTM

BiLSTM (Brueckner and Schulter, 2014) stands for bidirectional long short-term memory which is an extended version of recurrent neural network. BiLSTM employs two LSTM modules to distill inter and intra-relational structure from text using forward and backward feature learning strategy. In my proposed method, we employ the BiLSTM module with fused transformer features to overcome the shortfall of the transformer modules and extract the long-term causal relations from the text.

[1]https://huggingface.co/microsoft/deberta-v3-base

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

[3]https://huggingface.co/philschmid/distilroberta-base-ner-conll2003

Unsupervised contrastive learning achieve excellent success on different nlp tasks in recent times (Wang and Liu, 2021). We average each cause or effect span logit's logarithmic probability score using $log(Softmax(x))$ to calculate the loss.

2.3 Fusion Techniques

To enhance the performance of individual models or address model limitations, fusion is an effective strategy. In our proposed framework, we also employ two different types of fusion strategies for the proposed method of subtask 1 and subtask 2. For subtask 1, we employ late fusion, i.e. prediction level fusion, whereas in subtask 2 we employ early fusion strategy, i.e. feature level fusion. We utilized a weighted average of DeBERTa and Twitter-RoBERTa model predictions for late fusion where weights were 0.6 and 0.4 respectively.

3 Experiment and Evaluation

In this section, we now describe the dataset and hyper-parameters settings with the finetuning strategy that we have employed to design our proposed system for the CASE 2023 shared task 3.

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 2023 event causality shared task. The dataset for subtask 1 is same as CASE 2022 (Tan et al., 2022a) but this time subtask 2 dataset is enlarged as compared to the previous version, the current version extended 160 to 1981 sentences, 183 to 2754 causal relations in total.

3.2 Experimental Setting

We now describe the details of our experimental settings and the hyper-parameter settings with the fine-tuning strategy that we have employed to design our proposed CSECU-DSG system for the CASE 2023 event causality identification shared task. In our CSECU-DSG system, we utilize three state-of-the-art Huggingface transformer models with fine-tuning, including DeBERTa, Twitter-RoBERTa,and DistilRoBERTa. We use simple-transformers API (Rajapakse, 2019) to implement our proposed system for subtask 1. We use the train and development data during the model training phase. We used the CUDA-enabled GPU and

Subtask 1				
Team Name	F1 Score	Accuracy	Recall	Preision
DeepBlueAI (1)	0.8466	0.8466	0.8613	0.8324
rpatel12 (2)	0.8436	0.8409	0.8728	0.8162
timos (3)	0.8375	0.8324	0.8786	0.8000
CSECU-DSG (4)	0.8268	0.8239	0.8555	0.8000
elhammohammadi (5)	0.8245	0.8125	0.8960	0.7635
Subtask 2				
timos (1)	0.7279	-	0.6398	0.8442
tanfiona (2)	0.5971	-	0.5918	0.6025
CSECU-DSG (3)	0.3796	-	0.3612	0.4000

Table 2: Comparative performance with other selected participants. For subtasks 1 and 2 F1 scores denote binary and macro F1 scores, respectively.

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 and we used the default settings for the other parameters. In Subtask 2, we utilized augmented data provided by the organizer to train our model. The learning rate was 3e-05, batch size = 8, and we train the model for 10 epochs. The primary evaluation measure for both subtasks was the F1 score.

Method	F1 Score	Accuracy	Recall	Preision
CSECU-DSG	.8588	.8588	.8919	.8549
– Twitter-RoBERTa	.8470	.8470	.8756	.8481
– DeBERTa	.8538	.8538	.8972	.8469

Table 3: Individual component performance of our proposed method based on the development dataset of subtask 1.

3.3 Result and Analysis

The comparative results of our proposed CSECU-DSG system along with other top-performing systems (Tan et al., 2023) in subtasks 1 and 2 are presented in Table 2. Following the benchmark of CASE-2023 event causality identification subtask 1, participants' systems are ranked based on the primary evaluation metric F1 score where we see that our CSECU-DSG ranked 4th and 3rd in subtasks 1 and 2 respectively.

However, in subtask 1 our proposed system performance is relatively closer to top-performing systems which deduces the effectiveness of our system for causal event identification. On the other hand in subtask 2, though our system ranked well, still there are some limitations such that our model can predict only a single label for a single token whereas it may be a multi-labeled (cause or effect and signal) token which may hamper the system performance. In Table 3, we provide the individual component performance of our CSECU-DSG model of subtask 1. Where we can observe that the DeBERTa model is relatively well performed than the Twitter-RoBERTa which motivates us to employ a different transformer model in place of it for subtask 2.

4 Conclusion and Future Work

In this paper, we present two approaches to identifying causal events and extraction of causal relations from text. For the identification task, we proposed a unified neural network of two finetuned transformer models including DeBERTa and TwitterRoBERTa with a late-fusion technique. Similarly, for the extraction task, we utilize two transformers models but this time we incorporate the DistilRoBERTa model instead of the TwitterRoBERTa. Here, we design our model differently, we use DeBERTa with contrastive learning to train the cause-effect spans of text whereas DistilRoBERTa is used to train the signal span. Then we utilized an early fusion technique and pass the fused features to max-pooling and the final classifier label to get the predictions.

In the future, we intend to explore the challenges of nested causality extraction task where we will design a model to predict the multi-label of a single token at a time.

References

Ananth Balashankar, Sunandan Chakraborty, Samuel Fraiberger, and Lakshminarayanan Subramanian. 2019. Identifying predictive causal factors from news streams. 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 2338–2348.

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.

Raymond Brueckner and Björn Schulter. 2014. Social signal classification using deep blstm recurrent neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4823–4827. IEEE.

Quoc-Chinh Bui, Breanndán Ó Nualláin, Charles A Boucher, and Peter Sloot. 2010. Extracting causal relations on hiv drug resistance from literature. *BMC bioinformatics*, 11(1):1–11.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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.

Dominique Mariko, Hanna Abi Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (fincausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 105–107.

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

Jiangnan Qiu, Liwei Xu, Jie Zhai, and Ling Luo. 2017. Extracting causal relations from emergency cases based on conditional random fields. *Procedia computer science*, 112:1623–1632.

TC Rajapakse. 2019. Simple transformers. *URL: https://simpletransformers. ai/[accessed 2022-08-25]*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

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.