LTRC @ Causal News Corpus 2022: Extracting and Identifying Causal Elements using Adapters

Hiranmai Sri Adibhatla, Manish Shrivastava
Language Technologies Research Center, KCIS
IIIT Hyderabad, India
hiranmai.sri@research.iiit.ac.in,m.shrivastava@iiit.ac.in

Abstract

Causality detection and identification is centered on identifying semantic and cognitive connections in a sentence. In this paper, we describe the effort of team LTRC for Causal News Corpus - Event Causality Shared Task 2022 at the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022) (Tan et al., 2022a). The shared task consisted of two subtasks: 1) identifying if a sentence contains a causality relation, and 2) identifying spans of text that correspond to cause, effect and signals. We fine-tuned transformer-based models with adapters for both subtasks. Our best-performing models obtained a binary F1 score of 0.853 on held-out data for subtask 1 and a macro F1 score of 0.032 on held-out data for subtask 2. Our approach is ranked third in subtask 1 and fourth in subtask 2. The paper describes our experiments, solutions, and analysis in detail.

1 Introduction

A sizeable amount of text is generated every day due to increase in the amount of news available online from news portals and social media. Data available on social, political, and economics has the potential to revolutionise data-driven analysis (Barik et al., 2016). Causality identification and span detection (Do et al., 2011) is one such data-driven task. It is one of the many natural language processing (NLP) studies that attempts to address inference and comprehension. A causal relation is a semantic relationship between cause argument and effect argument such that the occurrence of one contributes to the occurrence of the other.

Cause is a span of text that results in the occurrence of an effect event. An effect is a span of text that is the consequence of the cause event and a signal is a span of text that binds both the cause and effect events. Together the study of cause and effect can help in understanding what agents contribute to the causes and the effects they create. Causality identification and span detection on climate science domain helps in analysing the rapid climate changes (Ionescu et al., 2020). Similarly analysis on financial domain news (Mariko et al., 2022) can help in improving trading strategies. Further examples include social, economic, and political sciences where the effects created by causes such as a change in policy can be identified over a period of time and analyzed.

Causal Text Mining have been shown to be beneficial for downstream tasks like summarization (Izumi et al., 2021; Hidey and McKeown, 2016), question answering and making inferences. Task 3 (Event causality identification) of CASE @ EMNLP 2022 (Tan et al., 2022a) aims at automatically identifying sentences that have a cause-effect event and extracting spans of text relating to cause, effect, and signal events. The shared Task 3 is divided into two sub-tasks:

Subtask 1: Causal Event Classification The first subtask identifies if a given event sentence contains any cause-effect.

Subtask 2: Cause-Effect-Signal Span Detection This subtask identifies the spans corresponding to cause and effect per sentence.

The causal news corpus (Tan et al., 2021, 2022b) comprises 3,559 event sentences, extracted from protest event news, that have been annotated with sequence labels on whether it contains causal relations or not. Subsequently, causal sentences are annotated with cause, effect, and signal spans. For both tasks, we use a Transformer-based model (Vaswani et al., 2017). We use adapters (Pfeiffer et al., 2020), a parameter-efficient fine-tuning method, in conjunction with a pre-trained model with strong language understanding and generation abilities (Liu et al., 2019). Recent research has shown that this method is robust to over-fitting in low-resource settings (He et al., 2021). In this way, the large pre-trained model RoBERTa, remains

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 50 - 55
December 7-8, 2022 ©2022 Association for Computational Linguistics

	Labels		
	Causal	Non-causal	Total
Train	1603	1322	2925
Dev	178	145	323
Test	176	135	311
Total	1975	1602	3559

Table 1: Data split for sentences in subtask 1

frozen, and only small modules the model parameters are optimized. This effectively retains acquired knowledge in the pre-trained language model. The first task was treated as a binary classification task with a single label for the input sentence, while for the second task, label was predicted for each input word of the sentence.

2 System Description

2.1 Data

The data consists of English news in the socio-political and crisis context, extracted from Automated Extraction of Socio-political Events from News (AESPEN) in 2020 (Hürriyetoğlu et al., 2020) and Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) in 2021 (Hürriyetoğlu et al., 2021) .

Figure 1 contains few annotated examples from the causal news corpus. The causes are highlighted in green, effects in purple and signals in cyan. Both cause and effect must be present in a same sentence to mark it as causal. The organizers made 3 datasets available for both the subtasks: *train, dev, and test*. Later UniCausal, a Causal Text Mining data (Tan et al., 2022c) was released to be used for both the subtasks. The labels for test data were not announced for both subtasks.

For subtask 1, around 869 news documents and 3559 English sentences were annotated with labels on whether they contained causal relations or not. Table 1 presents the sentence counts per data split.

For subtask 2, positive causal sentences from subtask 1 were retained and annotated with cause-effect-signal spans. From the total positive sentences, 180 sentences were annotated and there could be multiple relations per sentence. The data splits were: 130 train and 13 development.

After combining the causal news corpus and UniCausal corpus, the total number of unique samples on adding train and dev datasets are 6767 for subtask 1 and 1249 for subtask 2. We used 20% of the combined dataset for validation.

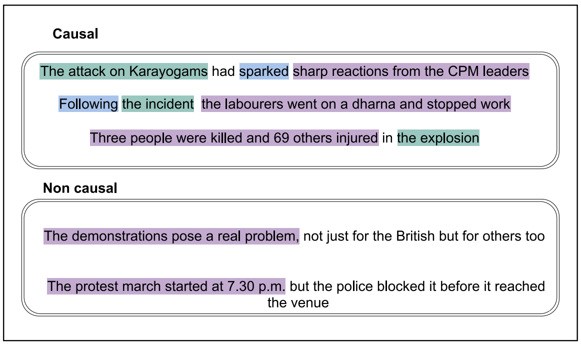

Figure 1: Annotated examples from Causal News Corpus. Causes are in highlighted in green, Effects in purple and Signals in cyan.

2.2 Solutions

Transformer based language models models (Vaswani et al., 2017) that have been pre-trained on massive amounts of text data and then fine-tuned on target tasks have resulted in significant advances in NLP (Liu, 2019; Yang et al., 2019), with state-of-the-art results across the board. However, models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have millions of parameters, making sharing and distributing fully fine-tuned models for each individual downstream task prohibitively expensive.

Adapters (Pfeiffer et al., 2020), which consist of only a small set of newly introduced parameters at each transformer layer, are a lightweight alternative to full model fine-tuning. Because of their modularity and compact size, adapters overcome several limitations associated with full model fine-tuning: they are parameter-efficient, they speed up training iterations, and they are shareable and composable. Furthermore, adapters typically outperform state-of-the-art full fine-tuning (Rücklé et al., 2020).

2.2.1 Subtask 1

Three transformers-based language models (Vaswani et al., 2017) were considered for the subtask 1 and fine-tuned on the causal news corpus dataset. The models experimented are BART (large) (Lewis et al., 2020), RoBERTa (base and large) (Delobelle et al., 2020) with an additional linear layer on top, RoBERTa (base and large) with adapter (Pfeiffer et al., 2020) and a classification head. Adapters are small learnt bottleneck layers inserted within each layer of a pre-trained model to avoid full fine-tuning of the entire model. The adapters framework enables them to be small, and scalable, particularly in low resource scenarios. It freezes all weights of the pre-trained model so only the adapter weights are updated during training. It activates the adapter and the prediction head such that both are used in every forward pass. As NLP tasks become more complex and necessitate knowledge that is not readily available in a pre-trained model (Ruder et al., 2019), adapters will provide a plethora of additional sources of relevant information that can be easily combined in a modular and efficient manner. We added a task-specific layer which is a classification head adapter. RoBERTa with classification adapter head and a linear layer added on top of RoBERTa (base) performed better than the BART-large model.

2.2.2 Subtask 2

Subtask 2 was modeled as a token classification task in the lines of named entity recognition (Li et al., 2020; Nadeau and Sekine, 2007) and parts-of-speech tagging (Schmid, 1994; Voutilainen, 2003). Each token of the cause effect sentence should be labeled as either cause, effect, signal or other. In the annotated data shared, span of text for cause was between ARG0 opening and closing tags, span of text for effect was between ARG1 closing and opening tags and span of text corresponding to signal enclosed between SIG0 opening and closing tags. The labeled annotations were pre-processed to be written in the Inside-Outside-Beginning (IOB) format (Ramshaw and Marcus, 1999) to aid in the identification of the sequences during inference. BertForTokenClassification model from BERT (base) (Devlin et al., 2019) was used for obtaining the contextual embeddings for the token and trained to predict the most probable label sequence. Since we saw a slight boost in performance on using adapters, we added a adapter head to RoBERTa (base) to predict the label sequence. In spite of using IOB format and contextual embeddings of BERT in modelling the problem as token labelling task, inference of predicted labels is difficult. A limitation that the model has is, that it can make an incorrect prediction in the middle of a cause/effect sequence or predict a cause/effect token in the middle of O tags. Few heuristics were employed to address the issue:

1. If a cause or effect sequence has a length lower than 2, it is ignored.

2. If a token is being preceded by a beginning-tag[1] and followed by either 'O' (for other) or the inside-tag [2], then the label is changed to its corresponding inside-tag.

3. If a token is predicted as (other) 'O', the sequence length of 'O' is less than 2, and is surrounded by beginning and inside tags of a single kind, then the label is changed to its corresponding inside-tag.

4. If a token is predicted as a beginning or inside tag of a kind, the whole sequence length is less than 2 and is surrounded by beginning and inside tags of another category, then the current category is changed to match the surrounding labels.

3 Evaluation

3.1 Experimental Setup

We fine-tune pre-trained transformers: BERT, BART and RoBERTa provided by huggingface [3]. The maximum sequence length for base models was 256 and for 512 for large model. The learning rate was 1e-4 and the models were fine tuned for 10 epochs for subtask 1 and 20 epochs for subtask 2. Adam optimizer was used with a dropout of 0.2 in each transformer layers. The train and validation batch sizes are 8 and 4 respectively.

3.2 Results

Model	R	P	F1
BART-large	0.85	0.81	0.84
RoBERTa-large+Adapter	0.82	0.84	0.83
RoBERTa-base+Adapter	**0.87**	**0.86**	**0.87**
RoBERTa-base+linear layer	0.86	0.83	0.84
Baseline	0.86	0.80	0.83

Table 2: Performance on Devset for subtask 1

Model	R	P	F1
BERT+Adapter	**0.056**	**0.023**	**0.032**
Baseline	0.003	0.009	0.005

Table 3: Performance on Devset for subtask 2

Table 2 shows the performance of our transformer based models for subtask 1 on the dev data

[1]The beginning-tags could be B-E for effect, B-C for cause and B-S for signal

[2]The inside-tags could be I-E for effect, I-C for cause and I-S for signal

[3]https://huggingface.co/docs/transformers

set. All the transformer variants have surpassed the baseline scores. RoBERTa (base) with adapters was our best-performing model. The slight improvement in precision and F1 scores for RoBERTa (base) with adapters over RoBERTa base with linear layer could be because, in the adapters framework, the adapters are added within each transformer layer while in the other approach, the linear layer is added to the output of the last layer of RoBERTa.

Table 3 shows the results obtained by using adapter on BERT (base). the predictions were post edited employing the heuristics discussed above. The results have improved marginally over the baseline model.

3.3 Error Analysis

While reviewing and analyzing the errors made by our models, we discovered few patterns where the models failed. Table 4 shows a few samples that were misclassified for subtask 1. We observed that the model fails to identify effects and causes that are not explicit. For the first example in Table 4, the effect is *"attracted a motley crowd"* and the cause *"the one-day fast"*. The cause phrase contains polysemous word *"fast"*, that could be misleading. In the second example *"raining bombs"* is a simile and in NLP tasks similies, idioms and proverbs have always been tough to comprehend. The model fails to identify phrases with length of less than four words without signal words. To check this further, we reordered the phrases in the second example and added a signal. The modified sentence we tested our model on was *"Mondal was hit by one of the bombs because both sides were raining bombs on each other, Murshidabad district magistrate Pervez Ahmed Siddiqui said"*. This sample does not change the meaning of the original sentence, but is reorganised and the conjunction is changed from a joining conjunction ('and') to a causal one ('because') and the model could classify the modified sentence as a causal sentence. False positives were also observed, the third and fourth examples contained an event or action, but the cause is not explicitly mentioned in the sentences. These incorrect predictions are a result of frequently encountering similar sentence structures in causal sentences. Longer sentences, having multiple clauses were also misclassified as causal sentences even when they are missing a cause of effect for the same reason.

Errors in subtask 2 were mainly because of incorrect and inconsistent predictions of cause and effect. The number of samples containing signals are very few in the dataset and therefore not well generalised by the model. As observed in Table 5, either the complete sequence is not predicted, or few tokens in the middle are incorrectly predicted.

4 Conclusion and Future Work

With the rapid growth in information from news portals, automated solutions to analyse data and draw inferences from the data play a pivotal role. Our solution for the both the subtasks involved adding an adapter layer which improves the performance by avoiding full fine-tuning of the entire model and instead adding additional newly initialized weights at every layer of the transformer which are trained during fine-tuning. Though the solutions work well, they could be further improved by using an ensemble model for subtask 1 and by adding an LSTM (Hochreiter and Schmidhuber, 1997) and CRF (Ye et al., 2009; Huang et al., 2015; Huang and Xu, 2015) on top of the contextual embeddings layer for proper alignment of tokens and labels for subtask 2.

In our experiments on the causal news corpus and on analyzing the misclassified samples we feel that the models for both subtasks can also benefit from having extra syntactic and semantic information. For subtask 1, verbs and signal arguments like conjunctions play a major role in determining if the sentence is causal or non-causal. Similarly for subtask 2, having part-of-speech tags information for all the tokens along with contextual embedding from BERT might work well. The current models have good contextual representations, but appending them with an extra embedding of the main verbs, conjunctions and parts-of-speech tags might steer the task inference in a better direction.

References

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

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Samples	Actual Label	Predicted Label
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
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
The police did not grant a permit for the march – the second time authorities have rejected a protest request – following a ban on the Saturday rally in Yuen Long	0	1

Table 4: Misclassified samples of subtask 1

Samples	Predicted Cause	Actual Cause	Predicted Effect	Actual Effect
The treating doctors said Sangram lost around lost around 5 kg due to the hunger strike.	due to the hunger strike	due to the hunger strike	The treating doctors said Sangram lost around 5 kg	Sangram lost around 5 kg
The Sadtu protest was a call for the resignation of Motshekga and her director general Bobby Soobrayan.	resignation of Motshekga	the resignation of Motshekga and her director general Bobby Soobrayan	The Sadtu protest was a call	The Sadtu protest
Troops also killed two militants making infiltration bids in Gurez sector today.	making infiltration bids in Gurez sector	making infiltration bids in Gurez sector today	Troops ('also' predicted as 'O') killed two militants	Troops also killed two militants

Table 5: Misclassified samples of subtask 2

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.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2208–2222, Online. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiheng Huang and Wei Xu. 2015. Kai yu. *Bidirectional LSTM-CRF models for sequence tagging. CoRR, abs/1508.01991*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Ali Hürriyetoğlu, Ali, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, and Erdem Yörük. 2021. 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, Ali, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020. Automated extraction of socio-political events from news (aespen): Workshop and shared task report. *arXiv preprint arXiv:2005.06070*.

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, Barcelona, Spain (Online). COLING.

Kiyoshi Izumi, Hitomi Sano, and Hiroki Sakaji. 2021. Economic causal-chain search and economic indicator prediction using textual data. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 19–25, Lancaster, United Kingdom. Association for Computational Linguistics.

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.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

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, 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.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2020. Adapterdrop: On the efficiency of adapters in transformers. *arXiv preprint arXiv:2010.11918*.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18.

Helmut Schmid. 1994. Part-of-speech tagging with neural networks. *arXiv preprint cmp-lg/9410018*.

Fiona Anting Tan, Devamanyu Hazarika, See-Kiong Ng, Soujanya Poria, and Roger Zimmermann. 2021. Causal augmentation for causal sentence classification. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 1–20, Punta Cana, Dominican Republic. 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 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 Thirteenth Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng. 2022c. Unicausal: Unified benchmark and model for causal text mining.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Atro Voutilainen. 2003. *Part-of-speech tagging*, volume 219. The Oxford handbook of computational linguistics.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.

Nan Ye, Wee Lee, Hai Chieu, and Dan Wu. 2009. Conditional random fields with high-order features for sequence labeling. *Advances in neural information processing systems*, 22.