SNU-Causality Lab @ Causal News Corpus 2022: Detecting Causality by Data Augmentation via Part-of-Speech tagging

Juhyeon Kim and **Yesong Choe** and **Sanghack Lee**
Graduate School of Data Science, Seoul National University, Seoul, South Korea, 08826
{kimjh9474,yesong,sanghack}@snu.ac.kr

Abstract

Finding causal relations in texts has been a challenge since it requires methods ranging from defining event ontologies to developing proper algorithmic approaches. In this paper, we developed a framework which classifies whether a given sentence contains a causal event. As our approach, we exploited an external corpus that has causal labels to overcome the small size of the original corpus (Causal News Corpus) provided by task organizers. Further, we employed a data augmentation technique utilizing Part-Of-Speech (POS) based on our observation that some parts of speech are more (or less) relevant to causality. Our approach especially improved the recall of detecting causal events in sentences.

1 Introduction

Nowadays, unprecedented amounts of data on social, political, and economic events offer a breakthrough potential for data-driven analytics. It drives and helps informed policy-making in the social and human sciences. Data of those humanities and social sciences cover a broad range of materials from structured numerical datasets to unstructured text data. An event is a specific occurrence of something that happens in a certain time and place involving humans. The events in texts can be understood in terms of causality, implies when one event, process, state, or object (namely, "cause") contributes to the production of another one (namely, "effect") where the cause is responsible for the effect.

Event-relating studies in the NLP have been growing, such as event extraction (EE), name entity recognition (NER), and relation extraction (RE). In particular, EE requires identifying the event, classifying event type and argument, and judging the argument role to collect knowledge about incidents found in texts (Li et al., 2021). Recent approaches to EE have taken advantage of dense features extractions by neural network models (Chen et al., 2015; Nguyen et al., 2016; Liu et al., 2018) as well as contextualized lexical representations from pre-trained language models (Wadden et al., 2019; Zhang et al., 2019).

However, there exist few studies regarding identifying or classifying events, especially based on causal relations. Phu and Nguyen (2021) studied Event Causality Identification (ECI) based on graph convolutional networks to learn document context-augmented representations for causality prediction between events. Cao et al. (2021) developed a model to learn a structure for event causality reasoning. Moreover, Man et al. (2022) introduced dependency path generation as a complementary task for ECI using causal label prediction.

In this study, we focus on causal event classification: whether a sentence contains any causal relation. Our framework employed both recent and traditional NLP techniques, which are pre-trained large language model (i.e., ELECTRA (Clark et al., 2020)) and POS tagging (Loper and Bird, 2002; Bird et al., 2009). To enhance the performance of detecting causality in each sentence, we attempted not only to concatenate another corpus that has causal labels but also to augment those corpora via POS tagging. With our base model, ELECTRA, those different combinations of datasets were compared to one another.

This paper is organized as follows. We first explore and examine the task and datasets. Based on the examination, we propose a new method in Section 3. We then present experimental results and discussion.

2 Task and Datasets

Causal event classification from natural language texts is a challenging open problem since causality in texts heavily relies on domain knowledge, which requires considerable human effort and time for annotating and feature engineering. In this study, as Subtask 1 of CASE-2022 Shared Task 3 (Tan

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 44 - 49
December 7-8, 2022 ©2022 Association for Computational Linguistics

et al., 2022a,b), we implemented causal event classification with large language pre-trained models.

The offered dataset is ‘Causal News Corpus (CNC)’ (Tan et al., 2022a). CNC contains sentences randomly sampled and refined from socio-political news. Each sentence in the dataset has a label, which represents whether it has a cause-effect relationship. This dataset was successfully used in Automated Extraction of Socio-political Events from News (AESPEN) at Language Resource and Evaluation Conference (LREC) 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). The number of training and validation data are 2925 and 323, respectively. Additionally, the organizers prepared the test set (which is only accessible through the task evaluation system) of size 311.

We additionally utilized an external dataset, ‘SemEval-2010,’ which was created for SemEval-2010 Task 8 (Hendrickx et al., 2019). The task was to classify semantic relations between pairs of nominals. One of the semantic relations is a causal relationship. Hence, we can directly infer whether a sentence in the dataset contains a causal relationship or not, allowing us to create another dataset to classify causality. `"The complication arose from the light irradiation."` is an example of a cause-effect labeled sentence from SemEval-2010. The training and test (used as validation) datasets contain 4450 and 786 sentences, respectively.

3 Methodology

CNC has a relatively small number of sentences to precisely detect whether any causal relation is contained in a sentence. Thus, we consider adding more sentences to CNC by (1) concatenating SemEval-2010 to CNC and (2) augmenting new sentences generated through POS tagging, which we will describe in the next section.

3.1 Data Augmentation via POS Tagging

A typical data augmentation is just attaching a new dataset to an existing original dataset. After augmentation, one may fine-tune the parameters of a model in hopes of improving performance of the model. Since a new dataset might come from a different distribution and features from the original one, it may negatively affect the performance. Hence, we propose to augment *causally relevant* information directly derived from the original datasets.

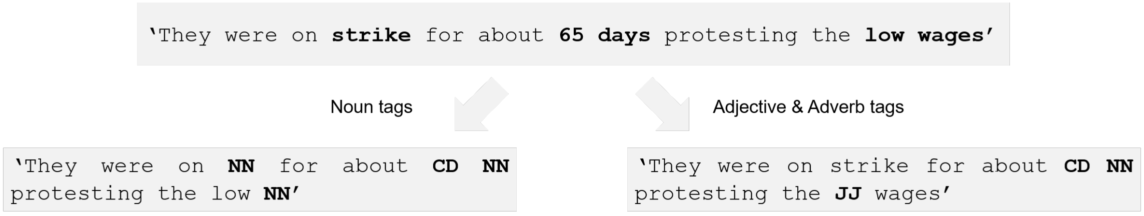
We argue that the causality in a sentence can be determined mainly by verbs and conjunctions, which is responsible for describing underlying causality, *not* nouns. That is, even if any nouns in a sentence are replaced with other nouns, a causal relation can still be preserved in the sentence. Consider `"There was a traffic jam as the taxi industry embarked on a protest"` for an example. Even if we eliminate the word `"traffic"`, the effect of `"protest"` is still `"jam"`. Regardless of the true meaning, there still exists a prominent causal relation. Hence, we proceed to exploit the following observation to devise our method: causal relationship is primarily captured by syntactic elements rather than semantics.

Against this background, we consider substituting words that are less likely to be related to causality (e.g., nouns, adjectives and adverbs) to their parts-of-speech, as depicted in Figure 1. This transformation preserves not only the original grammatical structure of the given sentence but also the underlying causality. Those newly transformed sentences were then concatenated to the original dataset for data augmentation.

One may consider replacing those words with any random words of the same POS as seen in *counterfactual augmentation* (Zmigrod et al., 2019). However, it could lead the model to learn wrong relationships since counterfactual sentences can cause spurious correlations with verbs or conjunctions. Thus, we just replaced those causally-irrelevant words with their corresponding POS tags.

3.2 Model

For our task, we initially considered three large pre-trained language models to construct a causal event classifier: Sentence-BERT, Span-BERT, and ELECTRA (ELECTRA-Base). We implemented the task with CNC for comparison among three models. Its result showed that ELECTRA outperformed other models. Therefore, we adopted our base model as ELECTRA. ELECTRA is trained via next sentence prediction similar to typical BERT models. Specifically, it learns through replaced token detection instead of masked language modeling. We conjecture that ELECTRA is effective especially for causal



'They were on **strike** for about **65 days** protesting the **low wages**'

'They were on **NN** for about **CD** **NN** protesting the low **NN**'

'They were on strike for about **CD** **NN** protesting the **JJ** wages'

Figure 1: Examples of POS tag-based sentences: 'NN' is a noun tag, 'JJ' is an adjective tag, 'RB' is an adverb tag, and 'CD' is a cardinal number tag. We have those transformed sentences added to the original dataset(s) to create new datasets (3), (4), (5) and (6).

detection since the causality in a sentence can be changed with just a single, crucial word change (i.e., replaced to a POS tag).

3.3 Experimental Setup

In this section, we explain various datasets used to train different ELECTRA models and hyperparameters to train the models. To utilize SemEval-2010, we pre-processed SemEval-2010 to make it similar to CNC—"label" is 1 if there exists causality in the sentence and 0 otherwise. To implement POS-tag based data augmentation, we used NLTK (Loper and Bird, 2002). We simply mention 'noun-base X' for X dataset with noun replaced to NN. We similarly define for adj/adv-base. We created six different augmented datasets:

1. CNC (2925 sentences)
2. CNC + SemEval-2010 (7375)
3. CNC + noun-base CNC (5850)
4. CNC + adj/adv-base CNC (5850)
5. CNC + SemEval-2010 + noun-base SemEval-2010 (11825)
6. CNC + SemEval-2010 + adj/adv-base SemEval-2010 (11825)

While we initially constructed other combinations of datasets, those six are interesting to compare and discuss. We used the following metrics accuracy, precision, recall and (Micro) F_1 score to measure the performance of trained models.

We used following hyperparameters to train ELECTRA models across the above six datasets.[1] The batch size is set to 32, and the epoch is set to 20. Gradient clipping is performed to prevent gradients from exploding, and the highest gradient is set to 1. In the beginning, the learning rate is set to 2e-5 so that it could learn in large steps. As the epoch iterates, the learning rate decreases with cosine annealing for the model to converge gradually. The optimizer used is *AdamW* (Loshchilov and Hutter, 2017) with a weight decay and a L_2 regularization added. Cross-entropy is used as a loss function. All models were neatly fit into a *single* NVIDIA Tesla V100 (16GB) GPU and trained efficiently and effectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Accuracy	0.849	0.841	0.855	0.849	0.852	**0.866**
Precision	0.865	0.865	0.838	0.838	0.865	**0.871**
Recall	0.871	0.859	**0.914**	0.901	0.882	0.908
F_1	0.866	0.862	0.874	0.868	0.874	**0.889**

Table 1: Performance of six models on the validation dataset where the models are trained on the datasets described in Section 3.3.

4 Results & Discussion

The performances of different datasets are compared (Table 1). Our results show that our proposed data augmentation method was effective.

4.1 Results

Our model trained on datasets with data augmentation achieved higher scores in all four measures. The recall increased remarkably: models with augmented datasets (3), (4) and (6) have the recall as 0.9 or above. While precision and recall are somewhat balanced across the models but precision is generally lower than recall. Due to the increase in recall, F_1 scores are all enhanced despite the increases in precision are negligible.

Compared to pure CNC (1), CNC with POS tag-base CNC (3, 4) produces better validation and test performances[2] than adding SemEval-2010 dataset (2) that also has causal labels but from a different distribution. Datasets (3) and (4) have recall above

[1] Our hyperparameters were not fully optimized in order to validate if our data augmentation method is effective so this is not for yielding the best of our model.

[2] Based on the performance reported in the leaderboard.

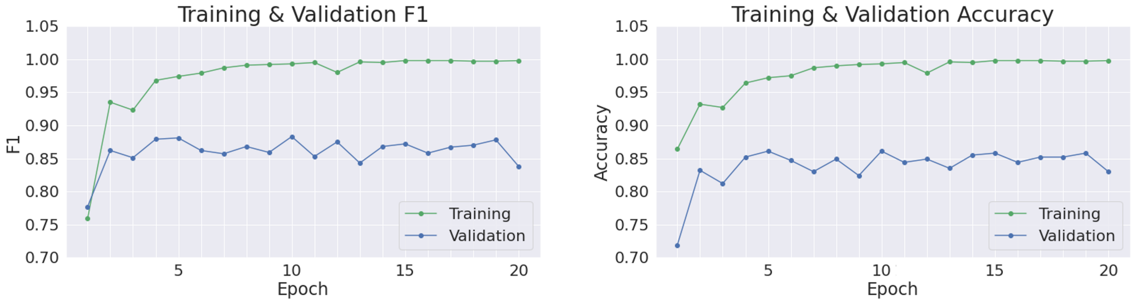


Figure 2: Training and validation F_1 scores (left) and accuracy (right) of dataset (6)

0.9, whereas dataset (2) has only 0.859.

Furthermore, dataset (6), which has SemEval-2010 and adj/adv-base SemEval-2010 added to the original CNC, achieved the highest F_1.

It is surprising given that adding SemEval-2010 *itself* (2) did not show improvements relative to (1). When it comes to the choice of POS tags to replace (noun (3, 5) vs. adj/adv (4, 6)), we do not have a consistent result to tell which tags are better to be replaced.

In Figure 2, we illustrate performance during training our model on (6). The accuracy and F_1 for the training dataset quickly reached 0.99 within 10 epochs in most of the experiments, and after it converges, the accuracies and F_1 scores were fluctuated slightly for the validation dataset.

Our model (6) was also evaluated with the test set through the task evaluation system. The model attained accuracy of 0.814, recall of 0.903, precision of 0.795, and F_1 of 0.848. The result is similar to what we have observed for the validation dataset.

4.2 Discussion

In this experiment, our model (6) trained with both SemEval-2010 and POS tag-base SemEval-2010 added to CNC attained the best performance in terms of accuracy and F_1 score. On account of the recall-precision trade-off, our results have higher recalls than precisions except for dataset (2). We think our model performs better with the sentences having causal relations since it seems to focus more on the features (e.g., embedding vectors) representing causality.

In the same vein, having a higher precision using the dataset with the SemEval-2010 added could be due to the more number of sentences having non-causal relations. Unlike other NLP corpora, not only the size of CNC is relatively small, but also there are not many causal-labeled datasets publicly available to additionally utilize. Furthermore, the ratio of the number of sentences that have causal relations to ones that do not is unbalanced (i.e., there is a way more number of sentences with no causal relations), so causal event classification is even more challenging. Thus, the data augmentation using POS tagging was effective and successful for this task. However, to increase the precision in the future, it is better to consider adjusting a threshold (i.e., decision boundary) for the results obtained through the current argmax function so that the model would not predict with certainty that causality exists when it truly did not.

We believe that our data augmentation method utilizing POS tagging can be generalizable and applicable to other learning methods. For instance, we found the benefit of the method for *prompt-based learning*, which allows the language model to be pre-trained on massive amounts of raw text to adapt to new scenarios with few or no labeled data (Liu et al., 2021). In our unreported experiment, we tried both original CNC sentences (i.e., dataset (1)) and their augmented one (i.e., dataset (3)) as prompt. Although both results were not as good as expected (i.e., the F_1 score is near 0.7), the result with having augmented dataset added had a higher recall, which corresponds to our results.

5 Conclusion

In this work, we proposed a framework that detects causal events from a sentence. In particular, because of the scarce number of sentences having causal relations, we devised a data augmentation strategy utilizing POS tags in place of causally irrelevant words. By augmenting the datasets, we indirectly increased the impact of verbs or conjunctions since causality relies on specific parts-of-speech in the context rather than the semantic meaning. The data augmentation strategy enhanced the performance of detecting causality especially in terms of recall and F_1. Given that the number of sentences having causal relations is small, detecting causality in those sentences is considered much more valuable than one in non-causal sentences.

Our contribution is that we provided an unconventional way of exploiting POS tags: previous studies using data augmentation via POS tagging *enhanced* the impact of specific words, such as informing proper nouns and word order for translation (Ding et al., 2020; Maimaiti et al., 2021). In contrast, we *weaken* the impact of specific words to indirectly improve the impact of other important words for detecting causality in sentences, such as verbs and conjunctions. By replacing those superfluous words with corresponding tags and adding those newly created sentences into the original corpus, our model outperformed those without data augmentation. This method can be a proper choice when adding new datasets is too expensive or there are few labeled datasets available.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. 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 4862–4872.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks. *arXiv preprint arXiv:2011.01549*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*.

Ali Hürriyetoğlu, 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, 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*.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, et al. 2021. A compact survey on event extraction: Approaches and applications. *arXiv e-prints*, pages arXiv–2107.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. *arXiv preprint arXiv:1809.09078*.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, Zegao Pan, and Maosong Sun. 2021. Improving data augmentation for low-resource nmt guided by pos-tagging and paraphrase embedding. *Transactions on Asian and Low-Resource Language Information Processing*, 20(6):1–21.

Hieu Man, Minh Nguyen, and Thien Nguyen. 2022. Event causality identification via generation of important context words. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 323–330.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 3480–3490.

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. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

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.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence*, 1(2):99–120.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.