Master’s Thesis of Data science

Detecting Causality by Data Augmentation via Part-of-Speech tagging

- De Novo Augmentation approach in NLP -

품사 활용 데이터 증강을 통한 인과관계 탐색

February 2023

Graduate School of Data Science
Seoul National University
Data Science Major
Juhyeon Kim

Detecting Causality by Data Augmentation via Part-of-Speech tagging

- De Novo Augmentation approach in NLP -

Sanghack Lee

Submitting a master's thesis of
Data Science

February 2023

Graduate School of Data Science
Seoul National University
Data Science Major

Juhyeon Kim

Confirming the master's thesis written by
Juhyeon Kim
February 2023

Chair ______________(Seal)
Vice Chair______________(Seal)
Examiner ______________(Seal)

Abstract

Keyword : Causal NLP, ELECTRA, Data Augmentation, POS tagging
Student Number : 2021-26348

With a deluge of text-based data available, the ability to automatically extract important information from the text data is crucial, especially extracting events from factual text data like news articles.

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, I developed a framework which classifies whether a given sentence contains a causal event.

As my approach, I 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, I employed a data augmentation technique utilizing Part-Of-Speech (POS) based on my observation that some parts of speech are more (or less) relevant to causality. The approach especially improved the recall of detecting causal events in sentences.

Chapter 1. Introduction

1.1. Study Background

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.

1.2. Purpose of Research

In this study, I focus on causal event classification: whether a sentence contains any causal relation. My 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, I attempted not only to concatenate another corpus that has causal labels but also to augment those corpora via POS tagging. With my base model, ELECTRA, those different combinations of datasets were compared to one another.

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

Chapter 2. Task and Dataset

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 (Tan et al., 2022a,b) of EMNLP (Empirical Methods in Natural Language Processing) 2022, I implemented causal event classification with large language pre-trained models.

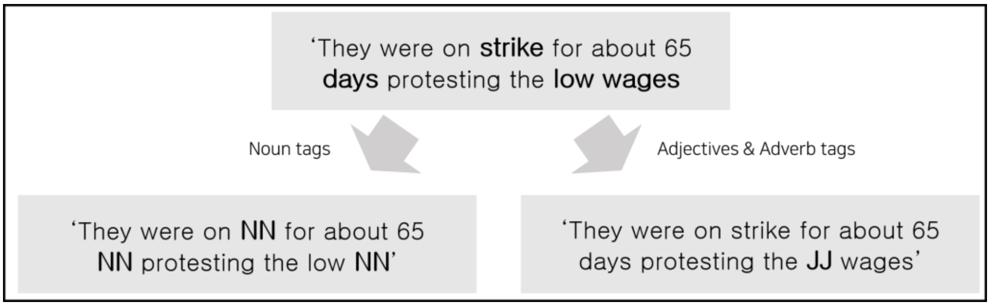
‘They were on strike for about 65 days protesting the low wages
Noun tags
Adjectives & Adverb tags
‘They were on NN for about 65 NN protesting the low NN’
‘They were on strike for about 65 days protesting the JJ wages’

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. I have those transformed sentences added to the original dataset(s) to create new datasets (3), (4), (5) and (6).

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 (Hürriyetoglu et al., 2020) and Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) in 2021 (Hürriyetoglu et al., 2021). The number of training and validation data are 2925 and 323, respectively. The organizers prepared the test set (which is only accessible through the task evaluation system) of size 311.

I 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, I 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. The longest sentence has 71 words, and the mean number of words in the sentence is 21 words.

Chapter 3. Methodology

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

3.1. Causal Graph of the task

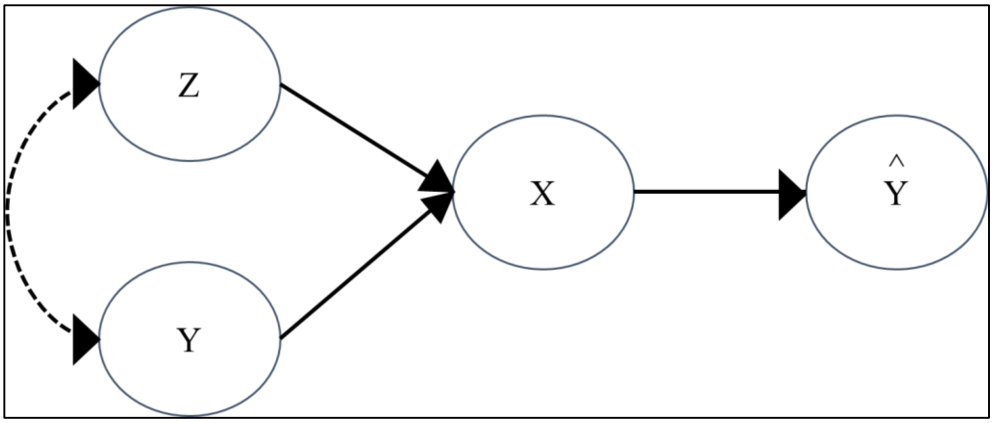

Figure 2. The causal graph (aka., Directed Acyclic Graph) of the task: X means text, Ŷ means label, Y means the component of the text, which is highly related to causality, and Z means the component of the text, which has spurious correlation with the text.

I drew a causal graph for the task. X means text itself, which is what I want to figure out whether there is causality or not. Ŷ means the label of the sentence, which contains causality or not: binary label. I assumed that there are the components that affect causality: verbs and conjunctions along with ones that do not affect: nouns, adjectives, and adverbs in the view of Part-of-speech. Especially, the components that are not related to causality have an important role of semantics but they do not have any effect on the existence of causality. In addition, there is an unobserved confounder between Y and Z because they may or may not have some kind of relation in grammar. For example, the form of verbs can be defined by the singular or plural of nouns. Hence, I attempted to have an intervention on Z to check Y's effect on Ŷ. This intervention is the data augmentation via POS tagging.

3.2. Data Augmentation via POS tagging

Since a new dataset might come from a different distribution and features from the original one, it may negatively affect the performance. Hence, I propose to augment *causally relevant* information directly derived from the original datasets.

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, I propose to augment causally relevant information directly derived from the original datasets.

Against the assumption I suggested, I 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 tags 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, I just replaced those causally-irrelevant words with their corresponding POS tags.

3.3. Model

There are three large language pretrained models used to perform the tasks: Sentence-BERT, Span-BERT, and ELECTRA. The reasons for using each model and a brief explanation are as follows.

3.3.1. SENTENCE-BERT

I expected that embedding each sentence with Sentence-BERT would have a great effect when considering my main task. Since the task does not require me to perform any detailed tagging for each element of a sentence like the NER task, to detect if there exists a causal relationship in each sentence with Sentence-BERT would be efficient. In addition, the biggest problem in the current dataset is the size. By calculating the similarity between sentences with Sentence-BERT, I can learn the representations of causal sentences using contrastive learning so that I can also implement data augmentation (Reimers et al., 2019).

3.3.2. SPAN-BERT

As Span-BERT randomly masks the token of the span, it would have excellent predictive power unlike any other general BERT models, which learn based on Masked Language Modeling and Next Sentence Prediction. Learning through Span Boundary Objective based on Masked Language Modeling with Span-BERT means that after learning by masking an arbitrary span of a token, it predicts a span at the boundary of the span. Therefore, it would be easy to find the span containing the causal relationship using Span-BERT (Joshi et al., 2020).

3.3.3. ELECTRA

ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) learns through Next Sentence Prediction like any normal BERT. Specifically, it learns through Replaced Token Detection instead of Masked Language Modeling. This is similar to Masked Language Modeling, but Replaced Token Detection learns in the form of determining whether or not it is an actual token by changing the masking target token to another token, rather than replacing it with a simple mask token (Clark et al., 2020). I expect that ELECTRA would be effective because causality can be changed with just one major word change. In the present experiment, the base model was used among the small, base, and large models.

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

I conjecture that ELECTRA is effective especially for causal detection since the causality in a sentence can be changed with just a single, crucial word change (i.e., replaced to a POS tag). In the present experiment, the base model was used among the small, base, and large models.

3.4. Experiment Setup

In this section, I explain various datasets used to train different ELECTRA models and hyper-parameters to train the models. To utilize SemEval-2010, I 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, I used NLTK (Loper and Bird, 2002). I simply mention ‘noun-base X’ for the X dataset with nouns replaced to NN. I similarly define for adj/adv-base. I 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 I initially constructed other combinations of datasets, those six are interesting to compare and discuss. I used the following metrics accuracy, precision, recall and (Micro) $F1$ score to measure the performance of trained models.

I used the following hyper-parameters to train ELECTRA models across the above six datasets. Those hyper-parameters were not fully optimized in order to validate if the data augmentation method is effective so this is not for yielding the best of my learning model.

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 $L2$ regularization added. All models were neatly fit into a *single* NVIDIA Tesla V100 (16GB) GPU and trained efficiently and effectively.

Chapter 4. Result and Discussion

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

	(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 Experimental setup.

4.1. Result

The 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.

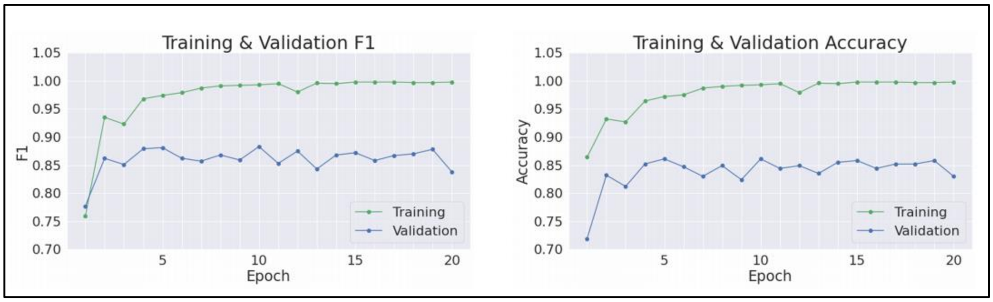

Figure 3: Training and validation *F*1 scores (left) and accuracy (right) of dataset (6)

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 (Based on the performance reported in the leaderboard of CASE @ EMNLP 2022 workshop) than adding SemEval-2010 dataset (2) that also has causal labels but from a different distribution. Datasets (3) and (4) have recall above 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 $F1$. 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)), I do not have a consistent result to tell which tags are better to be replaced.

In Figure 3, I illustrate performance during training the model on (6). The accuracy and $F1$ for the training dataset quickly reached 0.99 within 10 epochs in most of the experiments, and after it converges, the accuracies and $F1$ scores were fluctuated slightly for the validation dataset. Model (6) was also evaluated with the test set through the task evaluation system. The model obtained accuracy of 0.814, recall of 0.903, precision of 0.795, and $F1$ of 0.848. The result is similar to what I have observed for the validation dataset.

To find more interesting results, I experimented with other types of datasets. First, it is to make those datasets balanced for each class. Imbalanced datasets can decrease the performance of the model because a huge number of frequent class data can confuse the model with a tiny number of labels (Ramyachitra et al., 2014). There is the same issue as CNC and particularly for SemEval-2010. The ratio between sentences with causality and sentences without causality of CNC is 0.55. However, that of SemEval-2010 is 0.272. Thus, I changed those 6 datasets, which are described in section 3.4., to maintain the balance between classes so that each dataset has the same ratio of 0.5.

	(1)	(2)	(3)	(4)	(5)	(6)
Accuracy	0.726	0.711	0.701	0.765	0.705	0.709
Precision	0.708	0.678	0.743	0.754	0.730	0.718
Recall	0.885	0.912	0.686	0.856	0.746	0.765
$F1$	0.785	0.777	0.701	0.799	0.736	0.740

Table 2: Average performance of six models on the validation dataset where the models are trained on the datasets with balanced classes for 10 times with 10 different random seeds.

Overall scores decreased but the result shows a different tendency from the previous result (Table 2). (1) and (2) have higher recalls and get higher $F1$ scores than other augmented datasets except (4). The reason why this happens is that there are high deviations for the augmented datasets. For example, the range of recalls of (3) is from 0.55 to 1 without having a random seed fixed. Thus, there is enough information to learn even if those datasets are imbalanced, in addition to the fact that increasing the number of data leads to lower deviations in those scores.

	(1)	(2)	(3)	(4)	(5)	(6)
Accuracy	0.846	0.842	0.839	0.842	0.837	0.838
Precision	0.866	0.853	0.856	0.855	0.851	0.847
Recall	0.856	0.869	0.851	0.865	0.858	0.866
$F1$	0.860	0.860	0.853	0.859	0.854	0.856

Table 3: Average performance of six models on the validation dataset where the models are trained on the datasets with the same size of Table 2's dataset for 10 times with 10 different random seeds.

For the comparison, I also experimented with 6 datasets, which are the same sized ones as Table 2's datasets used but they are randomly sampled without any consideration for the ratio of class. Hence, their sizes of datasets are 3206, 5364, 6412, 6412, and 7522, respectively. The result shown in Table 3 indicates that the performance mostly gets much higher than Table 2's, but the overall performance is somewhat a little lower and different from the original results (Table 1). The reason is that experiments with the augmentations (3 to 6) reached accuracy 1 very quickly on each epoch even if the random seed changes every time so the model has rare chances to learn if sentences have causality or not.

I also experimented with the dataset for the test set. I transformed the test dataset into the POS tagged form because the model also learns the POS tagged form. I compare (3) and (4) for this experiment. The score was not as high as before. For (3), the accuracy, precision, recall, and $F1$ scores are 0.801, 0.829, 0.838, and 0.826, respectively. In addition, for (4), those scores are 0.828, 0.852, 0.853 and 0.846, correspondingly. Differences between precision and recall were decreased by transforming the test input.

4.2. Discussion

In this experiment, 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, the results have higher recalls than precisions except for dataset (2). I think the 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.

I believe that the data augmentation method utilizing POS tagging can be generalizable and applicable to other learning methods. For instance, I 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).

I 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 the results.

Chapter 5. Future work & Conclusion

5.1. Future work

This methodology is only specific to these datasets, task, and learning method. Therefore, I will need to improve the generalizability. For the generalization of this method, I will adapt and apply this method to 1) other datasets labeled with causality 2) other tasks related to causality in NLP, and 3) other learning methods such as prompt-based learning.

First, it is hard to find any dataset having causal relations due to the lack of annotators and the high labor cost. Moreover, it is hard to find any similar dataset with the task of CASE @ EMNLP 2022 because of the distinctive features of the ground truth sentences, which have causality included. Hence, I will transform Choice Of Plausible Alternatives (COPA) dataset similarly as I did with Sem-Eval 2010. COPA's task is to train a model to determine if a cause comes first before an effect, given two sentences with a binary label. For example, given two sentences: "`I am hungry`" and "`I had lunch,`" if the order of the input is "`I am hungry`" and "`I had lunch,`" the output should be `False`. On the other hand, if the input is opposite of the before one, the output should be `True`. Therefore, if I concatenate those two sentences that have `True` as their output, the dataset would be in the same format as the datasets used in this study.

Second, for the generalization of the task, I will adapt my methodology to other tasks related to causality such as COPA which is mentioned earlier. I will directly adapt the methodology to the COPA task and if possible, adapt it even in Korean as well. No matter which language it is as input, if my assumption holds, I believe the method I developed in this study would still perfectly work.

Third, for the generalization of the learning method, I will use prompt-based learning. Currently, prompt-based learning has been more popular, so if this augmentation approach can work in prompt-based learning, it can be generalizable above the learning techniques.

5.2. Conclusion

In this work, I proposed a framework that detects causal events from a sentence. In particular, because of the scarce number of sentences having causal relations, I devised a data augmentation strategy utilizing POS tags in place of causally irrelevant words. By augmenting the datasets, I 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 $F1$. 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.

The contribution is that I 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 (Maimaiti et al., 2021). In contrast, I 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 to the original corpus, the 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.

Bibliography

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Ali Hürriyeto˘glu, 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.

[8] Ali Hürriyeto˘glu, 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.

[9] 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.

[10] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, 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.

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

[12] Ilya Loshchilov and Frank Hutter. 2017. Decou- pled weight decay regularization. arXiv preprint arXiv:1711.05101.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Fiona Anting Tan, Ali Hürriyeto˘glu, 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.

[18] Fiona Anting Tan, Ali Hürriyeto˘glu, 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.

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

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

[21] 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.

[22] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

[23] Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, *8*, 64-77.

[24] Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, *5*(4), 1-29.