SPOCK @ Causal News Corpus 2022: Cause-Effect-Signal Span Detection Using Span-Based and Sequence Tagging Models

Anik Saha **Alex Gittens** **Bulent Yener**
Rensselaer Polytechnic Institute
{sahaa, gittea}@rpi.edu, yener@cs.rpi.edu

Oktie Hassanzadeh **Jian Ni** **Kavitha Srinivas**
IBM Research
{nij, hassanzadeh}@us.ibm.com, kavitha.srinivas@ibm.com

Abstract

Understanding causal relationship is an importance part of natural language processing. We address the causal information extraction problem with different neural models built on top of pre-trained transformer-based language models for identifying Cause, Effect and Signal spans, from news data sets. We use the Causal News Corpus subtask 2 training data set to train span-based and sequence tagging models. Our span-based model based on pre-trained BERT base weights achieves an F1 score of 47.48 on the test set with an accuracy score of 36.87 and obtained 3rd place in the Causal News Corpus 2022 shared task.

1 Introduction

Subtask 2 of the the Causal News Corpus shared task at the CASE-22 (Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text) addresses the causal information extraction problem (Tan et al., 2022). The goal of this task is to detect the spans of text in an input sentence that represent cause-effect pairs and, if extant, to also detect the text spans that "signal" this causal relationship. Figure 3 shows a sample from the data set. Simple examples of such signal spans are *result in*, *lead to*, and *due to*. In other cases, the causal relationship is implicit and it is important to understand the meaning of the whole sentence to detect causality. The Causal News Corpus data set contains sentences with both implicit and explicit causal relationship, so for this task, language understanding is an essential step. We adopt different pre-trained language models to develop our system owing to their tremendous success in natural language understanding tasks.

In this paper, we train and evaluate the performance of span-based and sequence tagging neural network models for the Cause-Effect-Signal Span Detection task. Our team name SPOCK (SPan and sequence based mOdels for Causal Knowledge) for the Causal News Corpus 2022 shared task is inspired by these model architectures. We trained a span-based (Eberts and Ulges, 2019) causality extraction system[1] by fine tuning the BERT-Base (Devlin et al., 2018) model. This model resulted in an F1 score of 47.48 and Accuracy score of 36.87. This was our best performing model compared to the ensemble of sequence tagging models based on the BIO scheme using the BERT-base and RoBERTa-large (Liu et al., 2019) language models.

2 Dataset and Task

We use the data sets from Causal News Corpus 2022 in our experiments. The sentences in this data set are collected from news sources containing event mentions. There are two subtasks in this challenge: subtask 1 is Causal Event Classification, where the goal is to determine if a sentence expresses a cause-effect relationship; subtask 2 is Cause, Effect and Signal Span Detection, where the goal is to identify the span of words in a sentence corresponding to a cause, effect, or signal (a span indicating the existence of a causal relation). This paper documents two approaches towards subtask 2. The training and dev set from subtask 2 are used for the training and evaluation of our models. In the final submission to the challenge, the trained models were used to obtain predictions on the test set.

This data set contains labels for Cause, Effect and Signal spans in a sentence whereas other commonly used data sets for causal relation extraction only contain labels for Cause and Effect. Further, it is possible for the Signal spans to overlap with the Cause or Effect spans. In some examples, the Signal words are not a contiguous span, i.e. words in different parts of the sentence are tagged as Signal. Data set statistics for subtask 2 are shown in Table 1.

[1] code for SpERT model available in https://github.com/aniksh/spert-causalnewscorpus

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 133 - 137

Data Split	Size
Train	180
Dev	323
Test	311

Table 1: Data set statistics

Each example in the training and dev sets is labeled with a single pair of Cause and Effect span. Some sentences contain multiple cause-effect pairs; each pair comprises a separate example, so that each example has a single cause and effect pair. Not all sentences in the data set contain a signal span. In some examples, the signal span overlaps with the cause or effect span. We show some examples in Figure 1.

3 Methodology

We experimented with two types of neural models for the Causal News Corpus 2022 challenge.

3.1 Span-based Model

We introduced this model in our submission (Saha et al., 2022) to the FinCausal 2022 challenge. The span-based model takes a sequence of tokens as input and predicts the Cause and Effect spans in the sentence by classifying a list of candidate spans of words. The list of candidate spans is generated by selecting all possible spans of words in the sentence up to a maximum span length. This model is based on SpERT (Eberts and Ulges, 2019) that classifies each span into 4 classes (Cause, Effect, Signal or None).

The input to the span classifier is a span embedding which takes the output layer embeddings from the *BERT-base* model. We split the words in a sentence with HuggingFace's *BertTokenizer* function (Wolf et al., 2019) to feed the pre-trained BERT model. We convert the annotations in the Causal News Corpus data set to Cause, Effect and Signal span labels for the span-based models.

The span-based model takes in a list of spans and builds an embedding for each span by using a max-pooling operation over the BERT output embeddings of the word pieces in that span. A context embedding is added to the span representation by concatenating the output layer embedding from BERT corresponding to the CLS token. The width of the span is included in the span representation by concatenating a span width embedding. The span-width embeddings are stored in a look-up table with a row for each unique span length of a cause or effect in the training data set. The embedding for a given span is thus the concatenation of the CLS token embedding, the width embedding, and a max-pool of the token embeddings in the span.

$$\mathbf{e}(s) = e_{CLS} \circ w_{k+1} \circ f(\mathbf{e}_i, \mathbf{e}_{i+1}, \ldots \mathbf{e}_{i+k})$$

where $\mathbf{e}(s)$ is the span embedding, e_{CLS} is the CLS token embedding, w_n is the width embedding for a span of size n and $\mathbf{e}_i$ the embedding for i-th token. A softmax layer is used on top of a linear classifier to convert the span embeddings into probabilities over 4 classes.

$$y_s = \text{softmax}(W_s \cdot \mathbf{e}(s) + b_s)$$

where W_s is the weight of the linear classifier and b_s is the bias of the linear classifier.

The cross-entropy loss is used to train the span classifier in this model. Spans are classified as either Cause, Effect, Signal, or None. Consider, for instance, the process of selecting a single Cause span. First we drop from consideration all spans whose probability of being a Cause are smaller than a threshold t. If there is no span left after applying the threshold, we predict there is no Cause in the sentence. Otherwise we take the Cause to be the span that achieves

$$\max_{s \in S} p_s$$

where S is the set of spans after dropping all spans below the threshold and all spans whose highest probability class is None, and p_s is the predicted probability for span s to be labeled as a Cause. Similar rules are used to identify the single Effect and Signal span.

Since the data set only contains positive labels for Cause, Effect and Signal spans, we generate negative examples by randomly sampling spans of words from the input sentence and labeling those as None. The negative span samples are selected from a list of all possible spans in the sentence up to the maximum span length from before. At inference time, a list of candidate spans is generated up to this maximum span size. We explain the span selection process in Appendix A. Since we predict Cause and Effect from a list of overlapping spans, the predicted Cause and Effect might possibly overlap but we did not face this problem as the span representation for overlapping spans are very similar.

<ARG1>Four students appeared in court on Monday</ARG1> <SIG0>for</SIG0> <ARG0>allegedly removing street signs</ARG0> .

Four	students	appeared	in	court	on	Monday	for	allegedly	removing	street	signs	.
B-E	I-E	I-E	I-E	I-E	I-E	I-E	O	B-C	I-C	I-C	I-C	O
O	O	O	O	O	O	O	B-S	O	O	O	O	O

<ARG1>The workers had embarked on a wildcat strike</ARG1> <ARG0><SIG0>demanding</SIG0> better working conditions</ARG0> .

The	workers	had	embarked	on	a	wildcat	strike	demanding	better	working	conditions	.
B-E	I-E	I-E	I-E	I-E	I-E	I-E	I-E	B-C	I-C	I-C	I-C	O
O	O	O	O	O	O	O	O	B-S	O	O	O	O

Figure 1: Examples with Cause, Effect and Signal span labels from the Causal News Corpus 2022 data set. The input text is labeled with ARG0, ARG1 and SIG0 labels. These are converted to the BIO tags for Cause-Effect and Signal as shown in different lines. The second example has overlapping Cause-Effect and Signal tags.

Model	Dev Set				Test Set			
	P	R	F1	Acc	P	R	F1	Acc
Baseline (Random)	2.17	2.17	2.17	20.84	0.30	0.89	0.45	21.94
Ensemble Tagging Model (BERT-base)	53.26	43.48	46.88	46.45	35.20	23.51	27.44	31.36
Ensemble Tagging Model (RoBERTa-large)	66.30	54.35	58.47	49.65	51.58	38.09	42.52	35.92
Span-based Model	56.52	72.16	62.62	44.71	57.62	43.75	47.48	36.87

Table 2: Precision (P), Recall (R), F1 and Accuracy score (Acc) of different sequence tagging models and the span-based model on the dev and test set.

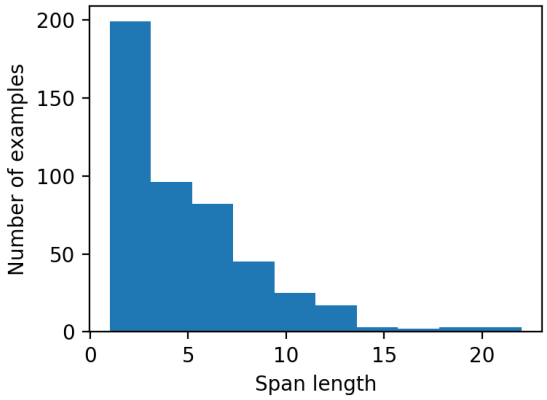

Figure 2: Span length distribution of the training set

The maximum span size is a hyperparameter for this model, chosen based on the distribution of the size of labeled Cause and Effect spans in the data set. Figure 2 plots the distribution of span sizes of all types in the training set. From our initial experiments, we found the 99-percentile span size from the training data to work well.

3.2 Sequence Tagging Models

This is a standard sequence tagging model that classifies each token in the sentence with BIO-style tags. The input text is tokenized with Huggingface tokenizers. For an input sequence, each token is assigned one of the following tags: {B-Cause, I-Cause, B-Effect, I-Effect, O}, where “B" stands for “Beginning", “I" for “Inside", and “O" for “Outside". Since this data set contains Signal span labels which overlap with the Cause and Effect labels we cannot represent these spans within a single sequence of BIO-style tags. To address the overlapping span problem, we introduce a separate set of tags for the Signal span. Figure 1 shows such an example with the BIO tags.

We experiment with both BERT-base and RoBERTa-large (Liu et al., 2019) as the encoder for the input sentence. The BERT-base model has 12 transformer layers with a token embedding dimension of 768 while the RoBERTa-large models has 24 layers with an embedding dimension of 1024. We add a 2-layer MLP to the output embeddings from the encoder to classify each token in the sentence. Since we have two sets of sequence tags, we train one MLP for detecting the Cause-Effect spans and another for detecting the Signal spans. These token classifiers share the same embedding representation. There are two cross-entropy loss functions for the two types of labels. We take a sum of these two loss functions as the total loss for the model. We fine-tune the pre-trained model weights and train the MLP parameters from scratch. We use the dev set performance to select the hyper-

parameters.

We take an ensemble approach to reduce the influence of randomness in the training on the final model performance. Specifically, we use majority voting to aggregate the token-level predictions on the test set from 11 different models trained with 11 different random seeds (0,10,20,... 100).

3.3 Training

We selected the hyperparameters by using the dev set performance as validation score and selecting the model with the highest F1 score. All models described here were trained on NVIDIA Tesla V100 gpus. We set the maximum span size to 20 as it covers 99% of the training data spans. The models are trained for 40 epochs with a learning rate of $5e^{-5}$. The number of negative samples per true label for the span classifier is set to 10.

4 Results

Span-based Model

The span-based model has a multi-class span classifier that predicts a score for each of the 4 classes. During inference, we filter all spans classified as None i.e. not a Cause, Effect or Signal. We assume the test data set might contain both causal and non-causal sentences, so we use a threshold ($t = 0.3$) on the predicted probability to filter spans which belong to a specific class (Cause or Effect). After thresholding, we select the span with the highest probability in each class.

This model achieves an F1 score of 47.48 and an Accuracy score of 36.87; it places 3rd in the shared task in terms of F1 score. It has the highest precision (57.62) among the submitted systems but low recall (43.75) value. We believe this model can achieve a higher score if we add a mechanism to predict multiple cause-effect pairs instead of a single cause-effect pairs.

Sequence Tagging Model

The sequence tagging model predicts both Cause-Effect and Signal tags to address the cases where these spans overlap. Since the model has a token-level classifier, it is possible that the predicted tags can form multiple spans for the same class. To convert the predicted token tags to span predictions, we take the first sequence of tokens in the sentence tagged in a class to be the single span for that class. We utilize only the class prediction to form the spans; in particular, either the 'B' or 'I' tags signals the start of a predicted span. The span prediction ends when the model predicts a different class for the next token or the sentence ends. We apply majority voting on the tags predicted for each tokens over 11 models trained with different random seeds. The ensemble method helps to reduce errors but we do not add any constraints to predict consecutive tokens. The RoBERTa-large model has 12% higher F1 score compared to the BERT-base model but it is lower than the Span-based model by about 4%.

Model	Text
Ground Truth	The treating doctors said Sangram lost around 5 kg due to the hunger strike .
BERT-base (Ensemble)	The treating doctors said Sangram lost around 5 kg due to the hunger strike .
RoBERTa-large (Ensemble)	The treating doctors said Sangram lost around 5 kg due to the hunger strike .
Span-based Model	The treating doctors said Sangram lost around 5 kg due to the hunger strike .

Figure 3: Sample predictions from the span-based model and the sequence tagging model. Yellow for Cause, Cyan for Effect, Red for Signal

.

Sample Prediction

We show the predictions from the sequence tagging and span-based models for the same input sentence in Figure 3. All 3 models label the same words as the Signal and the Cause spans. The BERT-base model predicts the wrong Effect span by selecting the phrase "treating doctors". The Cause-Effect span predictions from the RoBERTa-large model and the span-based models are the same. Since this sentence has a simple structure, it is relatively easier for these neural models to predict the Cause, Effect and Signal spans. The similarity in predictions from the span-based and the RoBERTa-large is also reflected in the results in Table 2 where these models have a small difference in F1 score.

5 Conclusion

In this paper, we adopt two approaches towards solving the Cause-Effect-Signal Detection task for participating in the subtask 2 of the Causal News

Corpus 2022 challenge. The span-based model outperforms the ensemble of sequence tagging models in both the dev set and the blind test set. In future work, we would like to adapt the models to predict multiple cause-effect pairs for a sentence. We will also focus on addressing the lack of large labeled data sets for this tasks by utilizing semi-supervised domain adaptation or generalization techniques.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. *arXiv preprint arXiv:1909.07755*.

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*.

Anik Saha, Jian Ni, Oktie Hassanzadeh, Alex Gittens, Kavitha Srinivas, and Bulent Yener. 2022. Spock at fincausal 2022: Causal information extraction using span-based and sequence tagging models. In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC*, pages 108–111.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Hansi Hettiarachchi, Tadashi Nomoto, Onur Uca, and Farhana Ferdousi Liza. 2022. 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.

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*.

A Span Selection

The span selection procedure is explained here with an example sentence. For the sentence, *The treating doctors said Sangram lost around 5 kg due to the hunger strike .* with a maximum span size of 5, we list all possible spans from size 1 to 5. We slide a window of a certain span size over the sentence to get all possible spans. For span size 3, the list of spans in this sentence would be - *[The, treating, doctors], [treating, doctors, said] . . . [hunger, strike, .].* So for each span size 1, 2, 3, 4, 5 we list all possible spans in the sentence to form the set of candidate spans.

Training. We select 10 negative samples randomly from each sentence during training. **Prediction.** To predict a Cause or Effect span, we need to list all possible spans from a sentence. So we classify all spans upto a maximum span size during inference.