Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 108–111
Marseille, 24 June 2022

SPOCK at FinCausal 2022: Causal Information Extraction Using Span-Based and Sequence Tagging Models

Anik Saha, Jian Ni, Oktie Hassanzadeh, Alex Gittens, Kavitha Srinivas, Bulent Yener
RPI, IBM Research, IBM Research, RPI, IBM Research, RPI
sahaa@rpi.edu, {nij, hassanzadeh}@us.ibm.com, gittea@rpi.edu, kavitha.srinivas@ibm.com, yener@cs.rpi.edu

Abstract

Causal information extraction is an important task in natural language processing, particularly in finance domain. In this work, we develop several information extraction models using pre-trained transformer-based language models for identifying cause and effect text spans from financial documents. We use FinCausal 2021 and 2022 data sets to train span-based and sequence tagging models. Our ensemble of sequence tagging models based on the RoBERTa-Large pre-trained language model achieves an F1 score of 94.70 with Exact Match score of 85.85 and obtains the 1st place in the FinCausal 2022 competition.

1. Introduction

An important step in extraction of causal information and narratives from text documents is the extraction of cause-effect pairs where causes and effects are text spans in the input sentences. The FinCausal shared task at the Financial Narrative Processing Workshop (FNP) addresses this step (Mariko et al., 2020). The causality information can be stated explicitly using well-known indicators such as *due to*, *caused by*, or *as a result of*. But in many cases, a causal relationship can be inferred based on the sequence of events even in the absence of specific patterns. This is more applicable to the financial domain where financial performance is often reported with the causal relation stated implicitly. Language understanding is an important step in extracting the cause-effect pairs from these financial reports.

In this paper, we address this information extraction problem with span-based and sequence tagging neural network models. Specifically, we fine tune pre-trained language models to perform text span classification and sequence labeling tasks. We trained a span-based (Eberts and Ulges, 2019) causality extraction system by fine tuning the BERT-Base (Devlin et al., 2018) model. This model resulted in an F1 score of 89.36 and Exact Match score of 81.67. Our best performing model was an ensemble of sequence tagging models based on the BIO scheme using the RoBERTa-Large (Liu et al., 2019) model. This model achieved an F1 score of 94.70 to win the FinCausal 2022 challenge.

2. System Description

We describe the two types of models trained for the FinCausal 2022 challenge.

2.1. Span-based Model

This model, based on (Eberts and Ulges, 2019), selects a sequence of tokens from the input text and classifies them to be a cause or an effect.

Preprocessing

We tokenize the text using the *word_tokenize* function from the `NLTK` library. To use BERT-Base model to get the embeddings, we split the tokens with the HuggingFace's *BertTokenizer* function (Wolf et al., 2019).

The FinCausal data set contains examples with multiple cause-effect pairs. These examples have the same input sentence with different cause and effect labels. There is an additional index number to denote these types of examples. Since our model takes the text as input, it is not possible for the model to predict two different labels for the same sentence. So we add a number to the start of these examples so the model has different inputs to work with. We follow the FinCausal 2020's winning system (Kao et al., 2020) to add a number to the start of the input text for multi-causal examples.

Model Description

We adopt the span-based model from (Eberts and Ulges, 2019) to classify spans of words as causes and effects. This model represents a span/sequence of words by max-pooling the output layer embeddings from BERT. The CLS token embedding is used as a context embedding in the span representation. The number of words in the span is embedded with a width embedding matrix to get a span width embedding. Span embedding is the concatenation of the span width embedding, max-pooled span embeddings and the CLS token embedding.

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

where $\mathbf{e}(s)$ is the span embedding, $\mathbf{e}_i$ the embedding for i-th token and w is the width embedding, c is the CLS token embedding. A candidate span is classified into 3 classes (cause, effect or none) using a softmax classifier.

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

There is also a binary relation classifier that is trained to predict the existence of a relationship between a pair of spans. The concatenation of the output embeddings from BERT and the max-pooled embeddings of the tokens in between the spans is used as input to the relation classifier.

This model is trained by selecting negative examples for the cause and effect spans by randomly sampling

1	Ceteris	paribus	,	the	fiscal	deficit	this	fiscal	will	widen	to	around
O	B-E	I-E	I-E	I-E	I-E	I-E	I-E	I-E	I-E	I-E	I-E	I-E
4	%	owing	to	the	stimulus	if	extra	transfers	from	RBI	are	counted
I-E	I-E	O	O	B-C	I-C	O	O	O	O	O	O	O
,	the	deficit	's	size	could	be	3.8	%	.			
O	O	O	O	O	O	O	O	O	O			

2	Ceteris	paribus	,	the	fiscal	deficit	this	fiscal	will	widen	to	around
O	O	O	O	O	O	O	O	O	O	O	O	O
4	%	owing	to	the	stimulus	if	extra	transfers	from	RBI	are	counted
O	O	O	O	O	O	O	B-C	I-C	I-C	I-C	I-C	I-C
,	the	deficit	's	size	could	be	3.8	%	.			
O	B-E	I-E	I-E	I-E	I-E	I-E	I-E	I-E	I-E			

Figure 1: Examples with multiple cause-effect pairs are distinguished by adding a number to the front. The BIO tags are shown under each token.

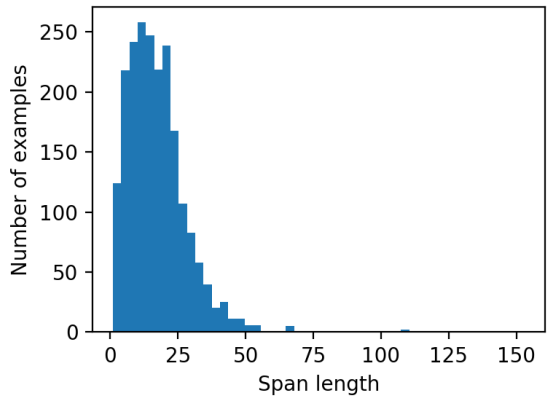

Figure 2: Span length distribution of the practice set

spans from the input text. We sample negative spans up to a maximum span size. During evaluation, a list of candidate spans is generated up to this maximum span length for predicting cause or effect with the span classifier. The cross-entropy loss function is used to train the model.

A challenge in the use of the span-based model is selecting the right span size. The distribution of the length of cause and effect spans in the training data is depicted in Figure 2. Our experiments showed that a span size equal to the 99-percentile span length (maximum span length after discarding the longest 1% spans) in the training data worked well across various data sets.

2.2. Sequence Tagging Models

This is a token classification model that predicts a tag for each token in the sentence using the output embeddings from RoBERTa-Large.

Preprocessing

We use NLTK and HuggingFace tokenizers for the input text. To format this problem as a token classification problem, we use BIO tagging scheme. 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". We also add a number at the start of examples with multiple cause-effect pairs. Figure 1 shows such an example with the BIO tags.

Model

We use RoBERTa-Large (Liu et al., 2019) as the input sequence encoder. This is a transformer model with 24 layers and the dimension of each layer embedding is 1024. A linear layer is added on top of the embeddings from the output layer to predict the BIO tags for each token. We fine-tune the model with the practice dataset and use the trial dataset for hyper-parameter tuning. For the final submissions, we submitted:

- Single models that are trained with practice data only.
- An ensemble model of 11 single models that are trained with practice data only via majority voting.
- Single models that are trained with all data.
- An ensemble of 11 single models that are trained with all data via majority voting.

3. Experiments

3.1. Data Set

We use the data sets from FinCausal 2021 in our experiments. The practice set is used as training set and the trial set is used as test set. For submission to the FinCausal 2022 challenge, we combine the practice set, trial set and additional practice set from FinCausal 2022 into a training set.

Model	Trial Set				Blind Test Set			
	F1	R	P	EM	F1	R	P	EM
BIO Tagging Model (Single) - Practice Data	87.92	88.00	87.98	75.63	94.57	94.57	94.62	83.06
BIO Tagging Model (Ensemble) - Practice Data	88.01	88.07	88.23	78.12	94.51	94.51	94.55	84.03
BIO Tagging Model (Single) - All Data	x	x	x	x	94.30	94.30	94.32	84.67
BIO Tagging Model (Ensemble) - All Data	x	x	x	x	**94.70**	**94.70**	**94.71**	**85.85**

Table 1: F1 score, Recall (R), Precision (P) and Exact Match score (EM) of different sequence tagging models on the trial and blind test sets.

Data Set	Size
Practice (FinCausal 2021)	1752
Trial (FinCausal 2021)	641
Practice-addition (FinCausal 2022)	535

Table 2: Data set statistics

3.2. Training

The span-based model was trained on a system with Tesla V100 gpu. We set the maximum span size to 60 as it covers 90% of the training data spans. The model is trained for 40 epochs with a learning rate of $5e-5$. The number of negative samples for the span classifier is 10. We selected the hyperparameters by using the trial set performance as validation score and selecting the model with highest score for exact matching.

4. Results

Span-based Model

The span-based model classifies candidate spans to predict cause and effect spans for a sentence. But it is possible that in some cases the model does not predict any cause or effect for an example. As we know, each example has 1 cause and effect pair in this data set, we modified the model prediction method. For each example we predict 1 span for the cause and effect classes by selecting the span with the maximum probability to be in the respective class. In Table 3, we see that predicting the span with the maximum probability to be a cause or an effect gives a big boost to the Exact Match score.

Model	F1	Rec.	Prec.	EM
SpERT	82.07	81.56	81.33	68.02
SpERT (Max)	83.94	83.57	83.42	74.10

Table 3: Result of the span-based model on the Trial data set

For submitting to the FinCausal challenge, we train this model by combining all data sets (practice, trial and practice-addition). This model gets a F1 score of 89.36 and Exact Match score of 81.67 (3rd rank in the competition in terms of Exact Match).

Sequence Tagging Model

The sequence tagging model based on RoBERTa-Large gets better partial F1 score compared to the span-based model. In the trial set, the single model trained on practice data achieves a F1 score 4% higher than the span-based model. We adopt the ensemble approach to improve the performance of this model. As random seeds play an important role in the optimization of deep networks, we train the same model with different random seeds and combine their prediction. We use majority voting as a simple approach to convert the predictions from different models into a single prediction. The ensemble approach ensures that the model does not have a low score due to a bad optimum resulting from a random seed. We submitted an ensemble of 11 models trained with different random seeds that obtains the best F1 score on the competition (Table 1).

Model	Text
SpERT (Max)	One of the pilot program's unique aspects is to encourage homeowners in six targeted community areas to opt in and put their houses in the land trust in exchange for significantly lower property taxes and access to a $30,000 grant for home repairs and energy upgrades.
BIO Tagging Model (Ensemble)	One of the pilot program's unique aspects is to encourage homeowners in six targeted community areas to opt in and put their houses in the land trust in exchange for significantly lower property taxes and access to a $30,000 grant for home repairs and energy upgrades.
SpERT (Max)	The group said international restaurant sales increased by 12.3 percent, benefiting from the opening of a record 20 restaurants during the year, but this was offset by a 15.9 percent sales decline in Australia and New Zealand.
BIO Tagging Model (Ensemble)	The group said international restaurant sales increased by 12.3 percent, benefiting from the opening of a record 20 restaurants during the year, but this was offset by a 15.9 percent sales decline in Australia and New Zealand.

Figure 3: Sample predictions from the span-based model and the sequence tagging model. for Cause and for Effect

Output Analysis

We compare the predictions from the span-based model and the sequence tagging model in Figure 3. The span-based model selects a shorter Cause phrase by focusing on the causal cue phrase 'to' whereas the sequence tagging model selects the clause before 'in exchange for' as the Cause phrase. In the second example, the predictions from the span-based model and the sequence tagging model are reverse, i.e. the span-based model classifies the first span as Cause but the sequence tagging model tags the first span as Effect. We can see that the sequence tagging model is correct here. As the span-based model predicts a span for each class, it can result to this type of error. So the sequence tagging model has a better performance on this task.

5. Conclusion

In this paper, we train different types of deep neural models based on pre-trained language models for the FinCausal 2022 shared task. We find that using an ensemble of sequence tagging models trained with the BIO tagging scheme based on the RoBERTa-large model achieves the best score in the competition.

6. References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eberts, M. and Ulges, A. (2019). Span-based joint entity and relation extraction with transformer pre-training. *arXiv preprint arXiv:1909.07755*.

Kao, P.-W., Chen, C.-C., Huang, H.-H., and Chen, H.-H. (2020). NTUNLPL at FinCausal 2020, task 2: Improving causality detection using Viterbi decoder. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 69–73, Barcelona, Spain (Online), December. COLING.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mariko, D., Labidurie, E., Ozturk, Y., Akl, H. A., and de Mazancourt, H. (2020). Data processing and annotation schemes for FinCausal shared task.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.