Domino at FinCausal 2020, Task 1 and 2: Causal Extraction System

Sharanya Chakravarthy* **Tushar Kanakagiri***
Karthik Radhakrishnan* **Anjana Umapathy***
Language Technologies Institute
Carnegie Mellon University
{sharanyc, tkanakag, kradhak2, aumapath}@cs.cmu.edu

Abstract

Automatic identification of cause-effect relationships from data is a challenging but important problem in artificial intelligence. Identifying semantic relationships has become increasingly important for multiple downstream applications like Question Answering, Information Retrieval and Event Prediction. In this work, we tackle the problem of causal relationship extraction from financial news using the FinCausal 2020 dataset. We tackle two tasks - 1) Detecting the presence of causal relationships and 2) Extracting segments corresponding to cause and effect from news snippets. We propose Transformer based sequence and token classification models with post-processing rules which achieve an F_1 score of 96.12 and 79.60 on Tasks 1 and 2 respectively.

1 Introduction

With the rise of the internet came an unprecedented growth in the amount of financial news being produced everyday. In its raw form however, this data though large, has limited utility. From this raw data, identifying relationship mapping between an external cause and its consequence can be of great value.

The simplest way to define a causal relationship is expressing it in the form of "E_1 causes E_2" where E_1 and E_2 are two linked events - E_1 being the cause and E_2 being the effect. Though this forms a simple illustrative example, real world causal relationships are often expressed with different lexical constructs, span across multiple sentences or are coupled together with other effects and causes.

We can now formally define our tasks as follows :

1. Binary classification task of identifying if a snippet of financial news displays a causal meaning. [Measured by weighted F_1]

2. Given a causal snippet, identify the span corresponding to the cause and the span corresponding to the effect displayed in the relation. [Measured by F_1 on span overlap]

Our first task is framed as a text classification task and has been widely studied in the NLP community (Minaee et al., 2020). Long-Short Term Memory networks, Convolutional Neural Networks (Kim, 2014), and more recently pre-trained Transformer models (Devlin et al., 2018) have shown success on a wide range of text classification tasks. Particularly, for the task of Cause-Effect detection, BERT based models have achieved state of the art performance (Soares et al., 2019) on the SemEval task (Hendrickx et al., 2019) on detecting semantic relations in text such as Cause-Effect, Entity-Origin, etc. To further leverage the benefits of pre-training, we use FinBERT (Araci, 2019), a BERT based model fine-tuned on financial data for better language understanding and domain adaptation.

Our second task is concerned with extracting spans corresponding to cause and effect from a causal news snippet. With the advent of large scale Question Answering datasets (Rajpurkar et al., 2016), deep learning models and specifically Transformers have been used extensively to identify answer spans. This task can also be framed as a token classification task where each token is classified as Cause/Effect/Other

*Equal contribution

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 90–94
Barcelona, Spain (Online), December 12, 2020.

and the tokens are grouped to form Cause/Effect spans. Additionally, token classification models such as BiLSTM-CRF (Li et al., 2019) and feature-based CRF (Mariko et al., 2020) have demonstrated good performance for causality extraction. We cast our task as a token classification problem and apply certain post-processing rules to reconstruct the spans from individual token outputs.

In this work, we present our results on Task 1 using a variant of BERT - FinBERT pre-trained on financial data. Our FinBERT model achieve F_1 scores of 95.60 and 96.12 on the validation and test set respectively.

For Task 2, we use BERT with a linear layer over token embeddings for token classification. Our model achieves a validation F_1 score of 75.40, beating the baseline CRF model and a test set F_1 score 79.60. The code for our approaches will be made available on GitHub[1].

2 Data

In this work, we use the two datasets provided by FinCausal 2020 Workshop (Mariko et al., 2020) to train and evaluate our models. In comparison to datasets for other NLP tasks, the datasets for Tasks 1 and 2 are relatively small.

2.1 Task 1

The Task 1 dataset consists of approximately 20,000 financial text samples, each annotated for the binary classification task of causality detection. The labels 1 and 0 indicate the presence and absence of a causal relationship respectively. The dataset has a large imbalance with only 7% of the examples labeled as 1.

2.2 Task 2

The dataset for Task 2 consists of 1,100 pieces of text containing causal relations, along with information about sections of the text that correspond to cause and effect. For example -
Text: Zhao found himself 60 million yuan indebted after losing 9,000 BTC in a single day
Cause: losing 9,000 BTC in a single day
Effect: Zhao found himself 60 million yuan indebted

Note that the system is expected to pick maximal spans - In the example shown above, the effect is annotated as "losing 9,000 BTC in a **single day**" although the fact that he lost it in a single day is irrelevant to the effect. Similarly for samples which span multiple sentences, the system is sometimes required to tag whole sentences as cause / effect even if the actual cause / effect is only a part of the sentence. The various rules and heuristics followed by the annotators are documented here [2].

3 Proposed Approach

We now present our systems for cause-effect detection and extraction. Given the small size of our dataset, it might not be possible for deep learning models to completely fit our tasks. We combat this in two ways - pre-training and post-processing. We leverage pre-training by utilizing FinBERT (Araci, 2019) - a BERT model trained on Reuters TRC2 Corpus, Financial Phrasebank and FIQA datasets for easier adaptation to the financial domain and use a post-processing module for Task 2 to incorporate some of the rules followed by the annotators as opposed to relying solely on the model to learn them.

3.1 Task 1

We first tokenize our sentence into sub-tokens (Wu et al., 2016) using the bert-base-uncased tokenizer provided by Hugging Face[3]. These tokens are passed through the FinBERT model[4]. The [CLS] representation from FinBERT is then passed through a 2 layer feed-forward network to output probability scores to indicate whether a cause-effect relation is contained in the sentence.

[1] https://github.com/sharanyarc96/Domino-Fincausal2020
[2] https://drive.google.com/drive/folders/1ryH76Z_hqrFaGlxx6SMJRGxK4BeJeg97
[3] https://huggingface.co/bert-base-uncased
[4] https://prosus-public.s3-eu-west-1.amazonaws.com/finbert/language-model/pytorch_model.bin

3.2 Task 2

We frame the task of extracting cause-effect spans as a token classification task by labeling each token as Cause/Effect/Other, and use a post-processing module to extract the spans from the predictions.

3.2.1 Pre-Processing

We first apply the bert-base-cased tokenizer[5] to our sentence and gold spans and tag each sub-token as Cause, Effect, and Other. For example,
Tokenized Sentence - "The stock market crashing resulted in multiple lay ##offs"
Labels - "C C C C O O E E E"

We manually corrected 11 instances in the training dataset where the gold Cause and Effect offsets were incorrectly labelled at the second letter of a word or before the end of a word, causing mismatches between the sentence and individual span tokenizations - an example of an incorrectly labelled gold cause is shown below.
Index - 0001.00010
Text - Connecticut, Pennsylvania, New Jersey, Illinois and New York lost about half of their income from people earning more than $200,000 - indicating the wealthy were picking up and leaving.
Gold Cause - *ndicating* the wealthy were picking up and leaving.
Effect - Connecticut, Pennsylvania, New Jersey, Illinois and New York lost about half of their income from people earning more than $200,000

3.2.2 Model

We use a softmax layer over each token representation produced by BERT to classify each token as Cause/Effect/Other. We then merge labels for sub-tokens and align them with the word tokenizations produced by NLTK[6] for compatibility with the evaluation script. For this task, we finetune the bert-base-cased model provided by Hugging Face[7].

3.2.3 Post-Processing Module

Since we framed the task as token classification, the model does not necessarily tag contiguous tokens with the same label. We apply rules based on adjacent tags to modify the tag in case of a mislabelling (A single cause token surrounded by effect tokens is likely mislabeled). Additionally, the task annotators followed various rules whilst tagging the sentences (marking entire sentences as cause/effect if the cause/effect were in different sentences, stretching the spans to cover the entire sentence, etc). We mimic these rules in our post-processing module to produce spans that better match the gold annotations.

Specifically, we check if the predicted tag is different from its previous and next, and if the previous and next tag are the same, we change the label of the predicted tag. In case of multiple cause-effect predictions, we take the dominant prediction (highest contiguous count) and if the cause and effects are predicted in individual sentences, we stretch their spans to encompass the whole sentence.

4 Experimental Setup

We finetune models for Tasks 1 and 2 on an NVIDIA Tesla T4 GPU and use the Adam optimizer (Kingma and Ba, 2014). The data partition and hyperparameters used for the individual tasks are described below.

4.1 Task 1

We finetune FinBERT for 8-10 epochs on the TRIAL set provided by the organizers, and report test set scores on the EVALUATION set. To finetune our hyperparameters, we train and evaluate on the data partition provided with the baseline code[8]. The test set in this partition is used to report validation set results. We experiment with batch sizes of 8, 16, 32 and learning rates in the range 5e-3 to 5e-5. The results reported are obtained using a batch size of 8 and a learning rate of 2e-5.

[5] https://huggingface.co/transformers/model_doc/bert.html#berttokenizer
[6] https://www.nltk.org/api/nltk.tokenize.html
[7] https://huggingface.co/bert-base-cased
[8] https://github.com/yseop/YseopLab/tree/develop/FNP_2020_FinCausal/baseline/task1

4.2 Task 2

We finetune BERT on TRIAL and use PRACTICE set as the validation set and report the test set scores on the EVALUATION set. We use a batch size of 8, learning rates in the range of 1.5e-4 to 5e-5 and run the optimization for around 15 epochs. We use the BERT tokenizer and pick the tag predicted for the root sub-token as the prediction for the word.

5 Results and Analysis

We compare our approaches against baselines and models provided by the task organizers. For Task 1, we compare our FinBERT model against a 1D CNN model (Kim, 2014) and vanilla BERT based task baseline. Tables 1, 3 show the strong performance of our model on validation and test splits. Our FinBERT model placed 7^{th} on the official FinCausal leaderboard[9].

For Task 2, we compare our approach against a zero-shot model based on SQuAD BERT[10] (We frame synthetic questions asking for the cause/effect, and extract spans produced by the model), a CRF baseline provided by the organizers and our Sequence Labelling approach. We can see from Table 2 that the post-processing module achieves gains of over 5 F_1 on validation. Our final model scores 79.60 on the official FinCausal leaderboard, placing 5^{th} on the shared task.

Model	F_1 **Score**
CNN-GloVe300D-FT	94.10
Task Baseline - BERT	95.23
FinBERT w/ fine-tuning	**95.60**

Table 1: Validation set results for Task 1

Model	F_1 **Score**
SQuAD BERT	43.22
Task Baseline - CRF	60.01
BERT Sequence Labelling	70.00
+ Post-Processing	**75.40**

Table 2: Validation set results for Task 2

Model	F_1 **Score**
FinBERT w/ fine-tuning	96.12

Table 3: Test set results for Task 1

Model	F_1 **Score**
BERT + Post-Processing	79.60

Table 4: Test set results for Task 2

5.1 Error Analysis

We randomly sampled from our predictions on Task 2 and analyzed the different kinds of errors made by the model. Errors primarily stemmed from the model's inability to correctly identify boundaries of causes and effects. Additionally, there were a few instances where the gold annotations had noise in tagging.

Some gold annotations appeared to have the cause and effect switched, such as the one shown below:
Index - 0038.00033
Text - CLIMATE - Invest in a 100% renewable electricity grid. Cut federal emission-level goals to 60% below 2005 levels by 2030.
Gold Cause - Cut federal emission-level goals to 60% below 2005 levels by 2030.
Error - Investing in renewable energy is the reasonable choice for the cause in this sentence and cutting emissions is more likely an effect and not the cause.

Some annotations had incorrectly labeled boundaries where the gold span started or ended within a word (§3.2.1). Despite correct prediction by our model, examples such as the one shown below are marked wrong due to switched cause-effect in the dataset.

The model incorrectly predicts the boundaries for causes and effects in multiple cases. On most occasions, the predictions are off by 1-2 words. To discern boundaries better, a larger training set could help the model make more accurate predictions or syntax-aware post-processing rules that identify phrases

[9] https://competitions.codalab.org/competitions/23748#results
[10] https://huggingface.co/distilbert-base-cased-distilled-squad

within sentences and ensure that the predicted spans cover entire phrases.
Index - 0047.00024
Text - Landmark 10 is racing four-year-old pacing mare, Hello Love, who races in the top condition events at Woodbine Mohawk Park and has made over $50,000 to date.
Predicted Cause - Landmark 10 is racing four-year-old pacing mare, Hello Love, who races in the top condition events at Woodbine Mohawk Park *and has*
Potential Solution - We can utilize the structural information by parsing the sentence and exclude 'and has' as they occur as part of a different sub-tree in the syntactic parse.

For a few examples with multiple sentences, the model sometimes includes irrelevant sentences as part of the cause/effect. Additionally, when a chain of cause-effect relations are present, our model predicts the final effect instead of the effect immediately following the cause. This increases the number of false positives and impacts the F_1 score. A potential solution would be to examine the individual logits in order to identify if strength of predictions vary significantly across the span and shorten the prediction in such cases.

6 Conclusion

In this work, we leverage BERT for sequence and token classification to detect and extract cause-effect relations in financial documents. We show the efficacy of domain-specific pre-training and address the low resource problem by applying post-processing rules over the model's predictions. In the future, we plan to experiment with data augmentation techniques such as synthetic data created from the cause and effect of different examples, and syntax-aware post-processing rules for more effective span extraction.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *ArXiv*, abs/1908.10063.

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*.

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*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2019. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *arXiv preprint arXiv:1904.07629*.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2020. The Financial Document Causality Detection Shared Task (FinCausal 2020). In *The 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP-FNS 2020*, Barcelona, Spain.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep learning based text classification: A comprehensive review. *arXiv preprint arXiv:2004.03705*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.