Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 116–120
Marseille, 24 June 2022

DCU-Lorcan at FinCausal 2022: Span-based Causality Extraction from Financial Documents using Pre-trained Language Models

Chenyang Lyu, Tianbo Ji, Quanwei Sun, Liting Zhou
School of Computing, Dublin City University
Dublin, Ireland
chenyang.lyu2@mail.dcu.ie, tianbo.ji2@mail.dcu.ie, quanwei.sun@insight-centre.org, liting.zhou@dcu.ie

Abstract

In this paper, we describe our DCU-Lorcan system for the FinCausal 2022 shared task: span-based cause and effect extraction from financial documents. We frame the FinCausal 2022 causality extraction task as a span extraction/sequence labeling task, our submitted systems are based on the contextualized word representations produced by pre-trained language models and linear layers predicting the label for each word, followed by post-processing heuristics. In experiments, we employ pre-trained language models including DistilBERT, BERT and SpanBERT. Our best performed system achieves F-1, Recall, Precision and Exact Match scores of 92.76, 92.77, 92.76 and 68.60 respectively. Additionally, we conduct experiments investigating the effect of data size to the performance of causality extraction model and an error analysis investigating the outputs in predictions.

Keywords: FinCausal 2022, span-based causality extraction, financial documents, pre-trained language models, sequence labeling

1. Introduction

The FinCausal 2022 shared task, as a part of the Financial Narrative Processing Workshop (El-Haj et al., 2020; El-Haj et al., 2021), aims to extract *cause* and *effect* from financial documents, where both *cause* and corresponding *effect* are spans in the original documents. Extracting causality spans from financial documents is not only important for causal understanding in financial texts but also helpful for improving natural language understanding in finance domain. FinCausal 2020 (Mariko et al., 2020) and FinCausal 2021 (Mariko et al., 2021) have established benchmarks for causality extraction task and significantly facilitated the development of methodology in this area.

In this work, we employ advanced pre-trained language models (PLMs) to facilitate the causality extraction task as PLMs have been proven to be effective in many NLP tasks including text classification, text generation especially on span extraction/sequence labeling task such as Named-entity Recognition and Question Answering (Devlin et al., 2019; Lewis et al., 2020; Qiu et al., 2020). Build on PLMs, we also propose a heuristically-induced *post-processing* strategy to refine the system predictions. Our best system (BERT-large + *post-process*) achieves F-1, Recall, Precision and Exact Match scores of 92.76, 92.77, 92.76 and 68.60 respectively. More importantly, we focus on investigating the effect of data size to the performance of causality extraction model in order to provide useful information for the development of methodology. We found that causality extraction models obtain fewer benefit from increasing data size when the training data contains more than 60% examples of the full training set. Additionally, we conduct analysis towards the errors occurred in the predictions of PLMs as well as in the annotations of the examples in the dataset.

2. Data

The data used in FinCausal 2022 task is created from Qwam [1] and Edgar database [2]. We show some examples in Table 1, moreover we show the data size and average length of *document*, *cause* and *effect* in each version of FinCausal in Table 2. From the average length in Table 2, we can see that FinCausal 2022 data has shorter *documents* and longer *cause* spans compared to early version of FinCausal. Therefore, that might pose new challenges for the FinCausal 2022 task. In FinCausal 2022, the employed data consists of data created in FinCausal 2020 and FinCausal 2021 as well as newly annotated data. The pre-processing steps in this work are listed as follows:

- For the training data used in this work, we combine the *practice* and *trial* data in early versions of FinCausal task and half of the newly annotated data provided in FinCausal 2022, we use the other half of the new data as dev set. After filtering, the resulting training data contains 4386 examples and the dev data has 265 examples.
- Its worth noting that one document can possibly contain more than one *cause-effect* pair, thus for the examples whose id ends with *'.1'* we prepend a *'First'* to their documents, and for the examples whose id ends with *'.2'* we prepend a *'Second'* to their documents, see the second and the thrid example in Table1.
- To tokenize the texts (*document, cause* and *effect*)in dataset, we employ the $word\_tokenize$ function in NLTK (Bird et al., 2009) [3].

[1] http://www.qwamci.com/
[2] https://www.sec.gov/edgar/search-and-access
[3] https://www.nltk.org

Document	Cause	Effect
Incumbent RBS boss Ross McEwan announced in April his intention to step down from his role at the head of the 62% state-owned banking giant, saying it was the right time to go having delivered on his strategy of stabilising the bank following its post-crisis bailout.	it was the right time to go having delivered on his strategy of stabilising the bank following its post-crisis bailout.	Incumbent RBS boss Ross McEwan announced in April his intention to step down from his role at the head of the 62% state-owned banking giant
First. Finally, ValuEngine cut shares of Gladstone Commercial from a buy rating to a hold rating in a research report on Monday, July 22nd. Three investment analysts have rated the stock with a hold rating and two have assigned a buy rating to the company's stock. The stock presently has an average rating of Hold and an average price target of $22.50.	Finally, ValuEngine cut shares of Gladstone Commercial from a buy rating to a hold rating in a research report on Monday, July 22nd.	The stock presently has an average rating of Hold and an average price target of $22.50.
Second. Finally, ValuEngine lowered shares of Travelers Companies from a buy rating to a hold rating in a research report on Thursday, August 1st. Two equities research analysts have rated the stock with a sell rating, ten have issued a hold rating and two have assigned a buy rating to the company's stock. The stock presently has a consensus rating of Hold and an average price target of $148.78.	Two equities research analysts have rated the stock with a sell rating, ten have issued a hold rating and two have assigned a buy rating to the company's stock.	The stock presently has a consensus rating of Hold and an average price target of $148.78.

Table 1: Examples of *document* and corresponding *cause* and *effect*, where the second and the third example have the sample input *document* but different *cause* and *effect* spans, thus we prepend a *First* to the second example and a *Second* to the third one in order tot enable the model to be able to distinguish them.

Dataset	Data Size	Document	Cause	Effect
FinCausal 2020	1750	50.11	20.57	20.57
FinCausal 2021	1752	49.79	20.64	20.27
FinCausal 2022	538	45.80	24.10	19.01
Overall Training	4386	49.87	20.70	20.26
Overall Dev	265	48.91	25.75	20.15

Table 2: The data size of examples and average length of *document*, *cause* and *effect* in FinCausal 2020, FinCausal 2021 and FinCausal 2022 and the training and dev set used in this work. For FinCausal 2020 and FinCausal 2021, the statistics are calculated based on the combination of the *practice* and *trial* data.

- For the label of each word, if a word is in *cause* span, then its label is *B-Cause* if it is the start of *cause* span otherwise its label is *I-Cause*, the same rule applies to the words in *effect* span. For the words outside of *cause* and *effect* span, we give them a *O* label.

3. Experiments

3.1. System

In this work, we employ advanced pre-trained language models including DistilBERT, BERT and SpanBERT. DistilBERT (Sanh et al., 2019) is the distilled version of BERT which is smaller and faster with a price of slightly lower performance, BERT (Devlin et al., 2019) is a powerful natural language understanding model which has been shown to be very effective on many NLP tasks and SpanBERT (Joshi et al., 2020) is an improved version of BERT, which adopts a specially-designed pre-training objective that predicts a continuous span in text, resulting in superior performance in span extraction tasks. On top of the the contextualized word representations produced by PLMs, we add extra linear layers to predict the probability that each word belongs to which label (*O, B-Cause, I-Cause, B-Effect, I-Effect*). During training process, the system is optimized using AdamW (Loshchilov and Hutter, 2019) with a *CrossEntropy* loss. In the inference time, we select the most probable (the label with the largest probability) label for each word and then decode the label sequence to corresponding *cause* and *effect* span. Based on the observations that our systems tend to predict spans that end with incomplete phrases or sentences, we proposed a simple post-processing strategy that heuristically removes the incomplete phrases and sentences in the tail.

3.2. Experiment Setup

We use the implementations of DistilBERT, BERT and SpanBERT from Huggingface (Wolf et al., 2020) [4]. The learning rate is set to 5e-5, weight decay rate is 0, we set the dropout rate to 0.1. We train our systems 30 epochs with a batch size of 16. All experiments are conducted on a NNIDIA GTX 3090 GPU.

3.3. Results

We show the main experimental results in Table 3, the systems we used include *DistilBERT, BERT-*

[4]https://huggingface.co/models

	Dev Set				Test Set			
	F-1	Recall	Precision	EM	F-1	Recall	Precision	EM
DistilBERT	87.31	85.69	89.79	0.015	89.21	89.21	89.22	17.26
+ post-procss	88.44	86.83	91.34	54.72	90.34	90.30	90.42	67.63
BERT-base	86.88	86.61	87.41	0.023	91.08	91.09	91.07	17.90
+ post-procss	88.10	87.78	88.75	56.23	92.23	92.20	92.28	68.70
BERT-large	91.40	91.40	91.42	0.015	91.60	91.65	91.64	18.11
+ post-procss	92.71	92.62	92.85	56.98	**92.76**	**92.77**	**92.76**	68.60
BERT-large-wwm	92.55	92.44	92.69	0.011	91.47	91.50	91.46	17.90
+ post-procss	93.87	93.67	94.25	58.11	92.61	92.60	92.62	**69.02**
SpanBERT-base	92.22	92.30	92.19	0.015	90.29	90.27	90.31	17.36
+ post-procss	93.59	93.57	93.62	59.25	91.44	91.38	91.55	67.95
SpanBERT-large	93.18	93.25	93.12	0.011	91.18	91.21	91.16	17.90
+ post-procss	**94.55**	**94.52**	**94.6**	**59.25**	92.35	92.34	92.36	68.70

Table 3: Experimental results of all systems on dev set and blind test set. Highest performance is in bold.

base, BERT-large, BERT-large-whole-word-masking[5]*, SpanBERT-base, SpanBERT-large*, we also show the effect of our proposed *post-process* strategy in Table 3. Our best performed system on blind test set (BERT-large + *post-process*) achieves F-1, Recall, Precise and Exact Match of 92.71, 92.62, 92.85, 56.9 on dev set and 92.76, 92.77, 92.76, 68.60 on the blind test set. The experimental results in Table 3 suggest:

- DistilBERT achieves comparable performance (slightly lower F-1, Recall and EM, higher Precision) with BERT-base while with a much smaller model size (40%$\times$Bert-base) and faster training and inference speed (50%$\times$Bert-base) compared to BERT-base, which is a huge advantage especially when deploying PLMs in production environment.
- For the same PLM, *large* model constantly yields performance better than *base* model. Moreover, the performance of PLMs is inline with their performance on other NLP tasks. For example, generally in terms of performance on NLP tasks: BERT-large-wwm $>$ BERT-large $>$ BERT-base, which is also true for FinCausal causality extraction task.
- The extremely low Exact Match score for all vanilla PLMs show that they struggle to precisely predict the correct boundary for the *cause* and *effect* spans in texts, suggesting that a vanilla PLM is still not enough for causality task although it can perform well on F-1, Recall and Precision scores.
- Our proposed *post-process* strategy substantially improve model's performance especially on Exact Match score. The results in Table 3 show that *post-process* can consistently give approximately 1.5 point improvements on F-1, Recall and Precision scores while significantly improve the Exact Match score. The results prove the effectiveness of our proposed *post-process* strategy.

[5]Referred to as BERT-large-wwm for simplicity

4. Analysis and Discussion

4.1. Effect of Data Size

We additionally conduct experiments investigating the effect of data size to the performance of causality extraction model. In experiment, we use increasing data sizes starting from 5% to 100% with intervals of 5%, we train our systems using the partial training data sampled from the full training set and evaluate all systems on the full dev set. For example, 5% training data means that we sample 5% examples from the full training set and use them to train a causality system and evaluate it on the dev set. The purpose of this experiment is to gain insights into how data size affects model's performance, in other words how much data is enough to yield a good performance. We show the curves of metrics (F-1, Recall, Precision and Exact Match) for the PLMs shown in Table 3 in Figure 1. The results show that all PLMs benefit from increasing data size at the early stage, however when data size exceeds 60% of the full training set (approximately 2600 examples) the performance has little improvements with increasing data size.

4.2. Error Analysis

We further analyse the errors in the predictions of PLMs, we randomly sampled some incorrect predictions from the output of *SpanBERT-large+post-process* and make manual analysis. The error type summarised from our manual analysis include:

- *Extra Content* (the predicted span contains more content than the golden one)
- *Less Content* (the predicted span contains fewer content than the golden one)

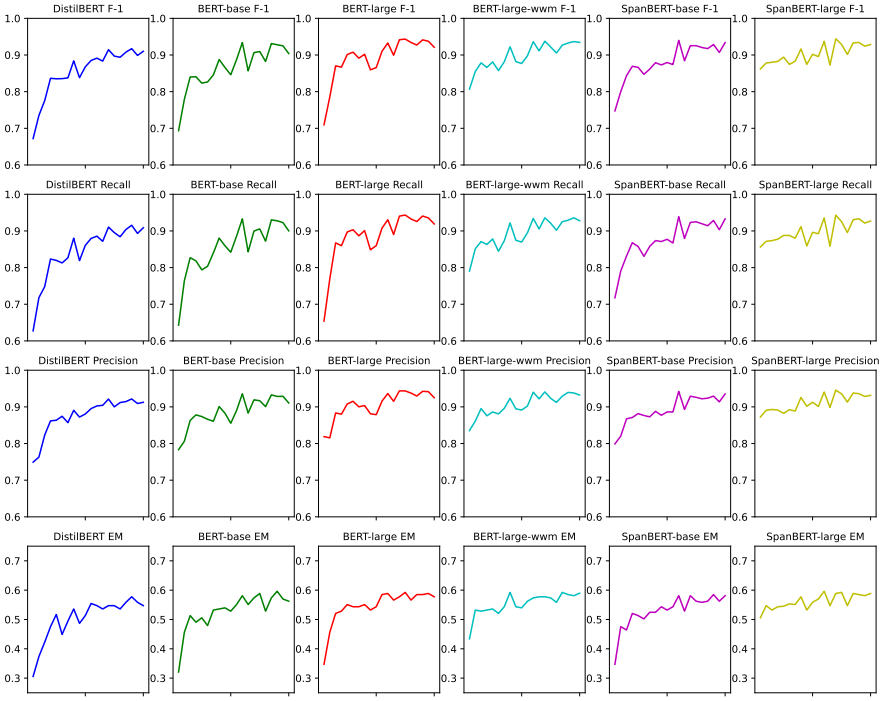

Figure 1: Visualization of metric curves of causality extraction models on different data sizes, where the y-axis is the metric score (F-1, Recall, Precision and Exact Match) and x-axis represents the data sizes starting from 5% to 100% with intervals of 5%.

Cause	Prediction	Error Type
the Company's Chief Executive Officer transition in 2011.	incremental costs associated with the Company's Chief Executive Officer transition in 2011.	*Extra Content*
Higher strategic SG&A costs in the technology businesses attributable to investments in strategic initiatives	Higher strategic SG&A costs in the technology businesses attributable to investments in strategic initiatives also	*Extra Content*
an after-tax charge of $305.1 million to settle certain patent litigation related to transcatheter mitral and tricuspid repair products.	settle certain patent litigation related to transcatheter mitral and tricuspid repair products.	*Less Content*
Working capital increased primarily due to the increase in accounts receivable and supplies inventory	Working capital increased	*Less Content*
lower incentive compensation costs in 2011 compared to 2010	lower incentive compensation costs in 2011 compared to 2010.	*Tail Punctuation*
Higher net charge-offs also contributed to the increase in the provision for credit losses and primarily reflect increases	as a result of the Merger.	*Completely Mismatch*

Table 4: Ground-truth *cause* span and corresponding prediction of *SpanBERT+post-process* associated with error type.

- *Tail Punctuation* (with an extra punctuation appended in the end of the predicted span)
- *Completely Mismatch* (completely different from the golden span)

We show some examples of incorrect predictions for *cause* spans in Table 4, these errors suggest that there is still room for improvements especially on Exact Match as both experiments results and error analysis show that PLMs have difficulty precisely predicting the boundary for *cause* and *effect* spans. Among all the errors, we think the *Tail Punctuation* is caused by the inconsitent annotation - if a ground-truth *cause* or *effect* span is a sentence or a clause including the end of a sentence or sub-sentence, it sometimes contains a punctuation (comma or full-stop) but sometimes it doesn't. That could cause confusion to the model in the training process, thus hindering the performance especially Exact Match score.

5. Bibliographical References

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dr Mahmoud El-Haj, et al., editors. (2020). *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, Barcelona, Spain (Online), December. COLING.

Mahmoud El-Haj, et al., editors. (2021). *Proceedings of the 3rd Financial Narrative Processing Workshop*, Lancaster, United Kingdom, 15-16 September. Association for Computational Linguistics.

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

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mariko, D., Abi Akl, H., Labidurie, E., Durfort, S., De Mazancourt, H., and El-Haj, M. (2020). The financial document causality detection shared task (fincausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32.

Mariko, D., Abi-Akl, H., Labidurie, E., Durfort, S., De Mazancourt, H., and El-Haj, M. (2021). The financial document causality detection shared task (fincausal 2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 58–60.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.