Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 135–138
Marseille, 24 June 2022

MNLP at FinCausal2022: Nested NER with a Generative Model

Jooyeon Lee **Luan Huy Pham** **Özlem Uzuner**
George Mason University
Fairfax, Virginia, USA
{jlee252,lpham6,ouzuner}@gmu.edu

Abstract

This paper describes work performed for the FinCasual 2022 Shared Task "Financial Document Causality Detection" (FinCausal 2022). As the name implies, the task involves extraction of casual and consequential elements from financial text. Our approach focuses employing Nested NER using the Text-to-Text Transformer (T5) (Raffel et al., 2020) generative transformer models while applying different combinations of datasets and tagging methods. Our system reports accuracy of 79% in Exact Match comparison and F-measure score of 92% token level measurement.

Keywords: Nest NER, Transformer Model, Generative Model, T5, NER

1. Introduction

In the field of financial analysis, the ability to swiftly and accurately comprehend the root causes and effects of events imparts valuable advantages in real-time decision making. The core obstacle to such a feat is the sheer volume and volatility of financial information which is being produced constantly. Our effort in this work is our contribution to this ongoing effort and research to address these challenges. The structure of the paper is simply: i) Methodology and Data, ii) Results and Discussion.

2. Methodology and Data

Our methodology generally leverages traditional generative systems. In a sequential manner, we started with text pre-processing, followed by fine-tuning the T5 model. Then, we employed post-processing to extract the correct span and entity. During the post-processing step, the system leverages specialized logic to select results among the output of multiple models to balance the strengths and weaknesses of each model in different scenarios.

2.1. Dataset Formulation

In addition to the official Fincausal dataset, we leveraged the Penn Discourse Treebank (PTDB) Version 3.0 Dataset (Miltsakaki et al., 2004). The third release of the PDTB, produced in 2020, contains data extracted from 2,499 stories from the Wall Street Journal over a three-year period, containing 53,676 tokens of annotated relations. It claims to be the largest such corpus of annotated relations available (Webber et al., 2019). We trained our model with different batches: 1) FNP only, 2) FNP and PDTB, 3) FNP and PDTB numeric values only, 4) FNP and PDTB Cause relations only 5) FNP and PDTB Result relations only 6) FNP and PDTB Implicit relations only 7) FNP and PDTB Explicit relations only.

2.1.1. FNP Dataset

The official Fincausal 2020 and 2022 dataset (FNP) of 2789 entries was extracted from a corpus of 2019 financial news as crawled and provided by Qwam. The official dataset, released and utilized since 2020, only includes entries with a 3-sentence distance between the cause and effect.

2.1.2. PDTB

The PDTB-style annotation uses a special pipeline-delimited format to identify spans of text and associated relationships. These relationships specified various forms of causal relations, identified as a subset of "Contingency Relations", where the "situation described by one argument provides the reason, explanation, or justification" (Webber et al., 2019) for the other. We extracted only the examples within the PDTB which resembled the cause-effect pattern, resulting in 7986 entries. For each cause-effect pair, we extracted the associated span of text which includes both members of the pair, as opposed to the entire full-length annotated article, thus maintaining consistency with the length of the official FNP dataset.

2.2. Pre-processing

To leverage the Generative Model, we created corresponding pairs of input and output and investigated the performance of different tagging methods. Examples of the raw dataset are shown in Table 1 and the tagged output in Table 2. We explored four methods. **Method 1** tags only the output, with the output including only the cause tags and effect entities, discarding all tokens which do reside in the entities. The effect span begins with a tag <e0> or <e1> and ends with a corresponding tag </e0> or </e1>. The cause span begins with a tag <c0> or <c1> and ends with the a corresponding tag </c0> or </c1>. **Method 2** is similar to Method 1, but retains the tokens outside of the cause and effect entities. **Method 3** involves tags on both input and output. Output is tagged using the same method as **Method 1**. For the input, we inserted a <causality> tag in front of the tokens indicating the causality, such as 'Due to', 'because', 'therefore', 'since', 'thus', 'if', 'as', 'when', 'after', 'as a result', 'subsequently', 'then', 'enhance', 'degrade', 'lead'. **Method 4** tags the output only, separating cause and effect with the tag <causality>, where a phrase before the tag is Cause, a phrase after the tag is Effect. This is the simplest method, though it does not consider nested cases. In this method, if two separate cause and effect pairs exist in one input, it is considered as two different inputs: one cause and effect pair is considered one input and second cause and effect are a separate, second input.

2.3. T5

One of the most challenging Natural Language Processing tasks is correctly recognizing named entities and their span, as they can partially overlap across different entities or also can be nested inside other entities altogether (Finkel and Manning, 2009). To address the issue of Nested NER, we use a generative model, T5 Transformer (Raffel et al., 2020), to generate the cause and effect from an input. In this paper, we compare the performance between models fine-tuned on a different dataset. We optimized hyper-parameters for each architecture using grid searches. This optimization includes varying learning rate (0.001, 0.002, 0.003, 0.0001, 0.0004 and 0.0005), batch size (8 and 16) and training epochs (2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28 and 30). We report results for each model using the hyper-parameters that yielded the highest accuracy. For all models, we set the max input length and output length to 200. For the generation step, beam size was set to 2 and repetition penalty was fixed to 2.5. All the experiments were conducted on the Google Colab Pro platform. The T5-base model available on Huggingface [1] [2] was used to fine-tune to our dataset.

2.4. Post-processing

We employed three sequential steps during the post-processing step. The first step is to correct common errors found during the validation testing. The second step is to extract the actual cause and effect using the cause tag (<c0>, <c1>, </c0>, </c1>) and effect tag (<e0>, <e1>, </e0>, </e1>). Finally, we select the best output from the different models.

- Step 1: Output Cleaning We have applied a cleaning process based on the validation output analysis that is described in the Section 3.2. The primary rule of cleaning are as follows: if there is any tag that is closed but not opened, then add the opening tag at the front of the entire output text.
- Step 2: Cause and Effect Extraction We simply extract cause by finding a phrase between open and close tag of cause, and effect by finding a phrase located between open and close tag of effect.
- Step 3: Model Selection We use ensemble learning techniques which shows higher accuracies in variety of tasks (Husain et al., 2020; Lee et al., 2021; Dang et al., 2020). Based on the validation accuracies, we selected the top 3 models: 1) Model trained with FNP only dataset with epoch 20 with learning rate 0.0001. 2) Model trained with FNP only dataset with epoch 28 with learning rate 0.0001 3) Model trained with FNP dataset and PDTB that contains numeric values with epoch 24 with learning rate of 0.0005.

3. Results and Discussion

3.1. Results

Accuracy is measured using an exact match of the gold standard string and generated strings for the validation. The validation sets are a randomly selected 20% portion of FNP data. The validation accuracy for data combinations are shown in Table 3a. The validation accuracy for different tagging methods are shown in Table 3b. The accuracy data shown in Table 3a are experiment results using tagging Method 1. The Table 3b has experiment results with dataset 1). The submitted output for the competition is a result of a model trained with both the training set and validation set we have.

3.2. Discussion

In this section, we show in depth error analysis to provide system implications for future development considerations. We focus on two different types of errors: tagging errors and span errors.

3.2.1. Common Tagging Errors

Case 1. Unclosed & unopened tags This is the case where a tag is opened, but never closed with the corresponding tag (i.e. <c0> exists, but </c0> not found).

- <c1> They set a sector perform rating and a $21.00 price target for the company. **</c0>** **<c0>** Seven equities research analysts have rated the stock with a hold rating and six have issued a buy rating to the stock. </c1> <e0> <e1> The company has an average rating of Hold and an average target price of $20.79. </e1> </e0>

Case 2. Cause and Effect Switched When cause is tagged as effect or effect is tagged as cause, it belongs to this case.

- Input: Consumer Banking and Wealth average total deposits increased $119.5 billion, or 120.4%, compared to 2019 driven primarily by the Merger and COVID-19 stimulus impacts.
- Gold Standard: <e0> Corporate and Commercial Banking average loans and leases held for investment increased $81.6 billion, or 95.9%, compared to 2019 </e0> <c0> the Merger and growth in corporate loans. </c0>
- Machine Output: <c0> Corporate and Commercial Banking average loans and leases held for investment increased $81.6 billion, or 95.9%, compared to 2019 </c0> <e0> the Merger and growth in corporate loans. <e0>

[1] https://huggingface.co/docs/transformers/model_doc/t5

[2] Our code is publicly available at https://github.com/jlee24282/FNP2022_MNLP

Index	**Text**	**Cause**	**Effect**
0009.00052.1	Things got worse when the Wall came down. GDP fell 20% between 1988 and 1993. There were suddenly hundreds of thousands of unemployed in a country that, under Communism, had had full employment.	Things got worse when the Wall came down.	GDP fell 20% between 1988 and 1993.
0009.00052.2	Things got worse when the Wall came down. GDP fell 20% between 1988 and 1993. There were suddenly hundreds of thousands of unemployed in a country that, under Communism, had had full employment. and PDTB	Things got worse when the Wall came down.	There were suddenly hundreds of thousands of unemployed in a country that, under Communism, had had full employment.
23.00006	In case where SGST refund is not applicable, the state is offering a 15% capital subsidy on investments made in Tamil Nadu till end of 2025	In case where SGST refund is not applicable	the state is offering a 15% capital subsidy on investments made in Tamil Nadu till end of 2025

Table 1: Three examples from FinCausal 2021 Corpus - Practice Dataset

	Input	**Output**
Method 1 (Single Relation)	Average short-term borrowings decreased as a percentage of funding sources due to strong deposit growth.	<e0> Average short-term borrowings decreased as a percentage of funding sources </e0> <c0> strong deposit growth. </c0>
Method 1 (Multiple Relations)	Things got worse when the Wall came down. GDP fell 20% between 1988 and 1993. There were suddenly hundreds of thousands of unemployed in a country that, under Communism, had had full employment.	<**c0**> <**c1**> Things got worse when the Wall came down.<**/c1**> <**/c0**> <**e0**> GDP fell 20% between 1988 and 1993. <**/e0**> <**e1**> There were suddenly hundreds of thousands of unemployed in a country that, under Communism, had had full employment. <**e1**>
Method 2 (Single Relation)	Average short-term borrowings decreased as a percentage of funding sources due to strong deposit growth.	<**e0**> Average short-term borrowings decreased as a percentage of funding sources <**/e0**> due to <**c0**> strong deposit growth. <**/c0**>
Method 2 (Multiple Relations)	Things got worse when the Wall came down. GDP fell 20% between 1988 and 1993. There were suddenly hundreds of thousands of unemployed in a country that, under Communism, had had full employment.	<**c0**> <**c1**> Things got worse when the Wall came down.<**/c1**> <**/c0**> <**e0**> GDP fell 20% between 1988 and 1993. <**/e0**> <**e1**> There were suddenly hundreds of thousands of unemployed in a country that, under Communism, had had full employment. <**e1**>
Method 3	<**causality**> In case where SGST refund is not applicable, the state is offering a 15% capital subsidy on investments made in Tamil Nadu till end of 2025	<**c0**> In case where SGST refund is not applicable <**/c0**> <**e0**> the state is offering a 15% capital subsidy on investments made in Tamil Nadu till end of 2025 <**/e0**>
Method 4	In case where SGST refund is not applicable, the state is offering a 15% capital subsidy on investments made in Tamil Nadu till end of 2025	In case where SGST refund is not applicable <**causality**> the state is offering a 15% capital subsidy on investments made in Tamil Nadu till end of 2025

Table 2: Examples of Nested NER format tagging from FinCausal 2021 Corpus Pre-processed.

Case 3. Incorrect Link between Cause and Effect We consider <c0> is a cause of <e0>, and <c1> is a cause of <e1>, while there should not be any link between (<c0> and <e0>) and (<c1> and <e1>). The following example shows a case where <e0> exists but not <c1>, <c1> exists not but not <e1>.

- Input: Consumer Banking and Wealth average total deposits increased $119.5 billion, or 120.4%, compared to 2019 driven primarily by the Merger and COVID-19 stimulus impacts.
- Gold Standard: <e0> Corporate and Commercial Banking average loans and leases held for investment increased $81.6 billion, or 95.9%, compared to 2019 </e0> <c0> the Merger and growth in corporate loans. </c0>
- Machine Output: <e0> Corporate and Commercial Banking average loans and leases held for investment increased $81.6 billion, or 95.9%, compared to 2019 </e0> <c1> the Merger and growth in corporate loans. </c1>

Case 4. Repetition When the tag meaninglessly repeats and causes an incorrect tag extraction, it belongs to this case.

- <c1> They set a sector perform rating and a $21.00 price target for the company.</c0> <**c0**> <**e0**> <**e0**> <**c0**> <**c0**> Seven equities research analysts have rated the stock with a hold rating and six have issued a buy rating to the stock. </c1> <**e0**> <**e0**>

	Dataset	Cause	Effect
1)	FNP	72.28	83.47
2)	FNP and PDTB	58.10	58.33
3)	FNP and PDTB numeric values only	68.60	69.53
4)	FNP and PDTB Cause relations only	67.9	67.33
5)	FNP and PDTB Result relations only	56.25	56.94
6)	FNP and PDTB Implicit relations only	53.01	49.5
7)	FNP and PDTB Explicit relations only	71.63	72.09

(a) Performance comparison between different dataset.

	Cause	Effect
Method 1	72.28	83.47
Method 2	69.60	74.53
Method 3	52.02	62.43
Method 4	66.89	65.21

(b) Performance comparison between different tagging method.

<**e0**> <e1> The company has an average rating of Hold and an average target price of $20.79. </e1> </e0>

3.2.2. Span Error

With the test output, we see that average exact match accuracy of the participants of Fincausal 2022 is 77.83%, while the F-measure score (measured at the token level) of 93.67%. This may be an indication that span errors are common among participants, given that considering relaxed match vs exact match increases accuracy by 14.8%. Our model shows the same tendency. Example of span error is as below.

- Input: Consumer Banking and Wealth average total deposits increased $119.5 billion, or 120.4%, compared to 2019 driven primarily by the Merger and COVID-19 stimulus impacts.
- Gold Standard: <e0> Consumer Banking and Wealth average total deposits increased $119.5 billion, or 120.4%, compared to 2019 </e0> <c0> the Merger and COVID-19 stimulus impacts.</c0>
- Machine Output: <e0> Consumer Banking and Wealth average total deposits increased 119.5 billion, or 120.4 <e0>, compared to <c0> 2019 driven primarily by the Merger and COVID-19 stimulus impacts.</c0>.

4. Conclusion

This paper shows a model submitted to FinCausal 2022 shared task as team MNLP. We studied different tagging methods and showed clear performance differences on the T5 generative model for the Nested NER task. We also explored the possibility of data amplification on the domain of financial cause and effect detection. The end result of our efforts culminated in a 79% Exact Match comparison score and a 92% F-measure score. With our experiments, we show the potential future directions with generative models for the Nest NER.

5. Bibliographical References

Dang, H., Lee, K., Henry, S., and Uzuner, Ö. (2020). Ensemble BERT for classifying medication-mentioning tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 37–41, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Finkel, J. R. and Manning, C. D. (2009). Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore, August. Association for Computational Linguistics.

Husain, F., Lee, J., Henry, S., and Uzuner, O. (2020). SalamNET at SemEval-2020 task 12: Deep learning approach for Arabic offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2133–2139, Barcelona (online), December. International Committee for Computational Linguistics.

Lee, J., Dang, H., Uzuner, O., and Henry, S. (2021). MNLP at MEDIQA 2021: Fine-tuning PEGASUS for consumer health question summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 320–327, Online, June. Association for Computational Linguistics.

Miltsakaki, E., Prasad, R., Joshi, A. K., and Webber, B. L. (2004). The penn discourse treebank. In *LREC*. Citeseer.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2019). The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.