NTUNLPL at FinCausal 2020, Task 2: Improving Causality Detection Using Viterbi Decoder

Pei-Wei Kao,[1] Chung-Chi Chen,[1] Hen-Hsen Huang,[2,3] Hsin-Hsi Chen[1,3]
[1] Department of Computer Science and Information Engineering
National Taiwan University, Taiwan
[2] Department of Computer Science, National Chengchi University, Taiwan
[3] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
{pwgao,cjchen}@nlg.csie.ntu.edu.tw,
hhhuang@nccu.edu.tw, hhchen@ntu.edu.tw

Abstract

In order to provide an explanation of machine learning models, causality detection attracts lots of attention in the artificial intelligence research community. In this paper, we explore the cause-effect detection in financial news and propose an approach, which combines the BIO scheme with the Viterbi decoder for addressing this challenge. Our approach is ranked the first in the official run of cause-effect detection (Task 2) of the FinCausal-2020 shared task. We not only report the implementation details and ablation analysis in this paper, but also publish our code for academic usage.

1 Introduction

Adopting causality information as features can benefit lots of applications such as question answering (Sharp et al., 2016), event prediction (Balashankar et al., 2019), and medical text mining (Khoo et al., 2000). In the financial domain, causality detection can be applied to stock movement prediction (Balashankar et al., 2019) and supporting financial services (Izumi and Sakaji, 2019). To better explain the causality that occurs between financial events, cause-effect detection is a fundamental research issue.

Taking a close look to financial documents, we find that there may exist multiple causal events and multiple causal chains in a paragraph. In such a case, traditional extraction methods like discourse parser are not feasible. In order to deal with this issue, we formulate the cause-effect detection task as a sequence labeling problem and propose an approach using BIO scheme and Viterbi decoder. We find that the proposed approach has the ability to identify multiple causal events and multiple causal chains in a given short paragraph, and also gives a better event span boundary.

Our contributions are two-fold as follows.

1. We propose an approach to cause-effect detection for financial news that could better identify multiple causal events and event spans.

2. We release the code of the best-performing model for future research.[1]

2 Pre-processing

We experiment on the FinCausal-2020 dataset (Mariko et al., 2020), which consists of two subtasks, including causal meanings detection (Task 1) and cause-effect detection (Task 2). The numbers of training instances are 22,058 and 1,750 for Task 1 and Task 2, respectively. This work only focuses on Task 2.

We use the Stanford CoreNLP Stanza (Manning et al., 2014; Qi et al., 2020) toolkit[2] to tokenize each sentence and generate the part-of-speech (POS) tag for each token. For the examples with multiple causal events, we recognize them by their indices and add a special number token before each example to treat them as different model inputs. As for causal relations tagging, we use "B, I, O" (Begin, Inside, and Outside) and "C, E" (Cause and Effect) labels to represent the positional information of the words and the semantic roles of the causal events.

[1] https://github.com/pxpxkao/FinCausal-2020

[2] https://github.com/stanfordnlp/stanza

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 69–73
Barcelona, Spain (Online), December 12, 2020.

3 Methods

3.1 Baseline Models

- **Conditional Random Field (CRF)**: CRF (Lafferty et al., 2001) is a popular model for sequential labeling, which considers the neighboring labels during calculation. We use the default parameter settings provided by Mariko et al. (2020) to train a baseline model for comparison.

- **Bidirectional Encoder Representations from Transformers (BERT)**: The pretrained text encoder BERT generally performs well in many NLP tasks (Devlin et al., 2018). In this paper, we use the *BERT-base* model, which consists of 12 Transformer layers with the hidden dimension of 768. It is pre-trained on *Masked Language Model Task* and *Next Sentence Prediction Task* via a large cross domain corpus and is well-known for its simplification of fine-tuning downstream tasks. We implement this baseline model by using the package provided by *huggingface*[3] (Wolf et al., 2019).

  For clarity of the model structure, we add a linear layer above the BERT-base model, and fine-tune it into a token classifier as our baseline for experiment. Given a sequence of tokens in the source documents $\mathbf{x} = [x_1, x_2, ..., x_n]$, the classifier will generate the target sequence of labels $\mathbf{y} = [y_1, y_2, ..., y_n]$, $y_i \in \{O, C, E\}$ for baseline target, and $y_i \in \{O, B-C, I-C, B-E, I-E\}$ for BIO scheme target.

  The training parameters of max length, batch size, and epoch is set as 350, 4 and 4, respectively. The initial learning rate is set to 5e-05, and we use cross entropy as the loss function. The final training time with one GTX TITAN X and Core i7-6700 is approximately 15 minutes, and requires 4GB GPU RAM.

3.2 POS Feature

With the POS tags labeled by the Stanford CoreNLP Stanza (Manning et al., 2014; Qi et al., 2020) toolkit, we first represent the POS information as a one-hot vector and concatenate it with the output of BERT's last hidden state output, and then send it to the final linear layer. The concatenated tensor is subject to predict the final label of the token.

3.3 Viterbi Decoder

Viterbi decoding is a dynamical programming algorithm that allows us to find the path with the global optimal probability. The probability of the most possible path ending in state t with observation i is:

$$p_t(i, x) = e_t(i) \max_k (p_k(j, x-1) \times p_{kt}), \tag{1}$$

where $e_t(i)$ represents the emission probability to observe element i in state t, $p_k(j, x-1)$ represents the probability of the most possible path ending at position $x-1$ in state k with element j, and p_{kt} represents the transition probability from state k to t.

Our proposed Viterbi decoder is added upon the fine-tuned BERT classifier only during evaluation. We consider the final output of the linear classifier in BERT as the emission matrix, and pre-define the transition matrix based on the BIO scheme. The Viterbi decoder will compute recursively to find the most probable sequence.

4 Results

We perform five-fold cross-validation to verify our experimental results during the self-evaluation period. Table 1 shows the results of the self-evaluation, where exact match stands for the accuracy of both predicted cause and effect are exactly matched with the labels. Table 2 shows the results of the blind test round. The baseline BERT-base model outperforms the CRF model by approximately 0.2 in terms of F1 score and 0.4 in terms of exact match ratio, showing the advantage of the pre-trained model. Compared with CRF, we notice that BERT can better distinguish multiple causal events and causal chains given the

[3] https://github.com/huggingface/transformers

same input text. We also find that adding POS features does not significantly improve the performance. Besides, using the BIO tagging scheme alone decreases the performance. However, combining the BIO tagging scheme with the Viterbi decoder achieves a great improvement in exact match ratio by rising 5%.

Extractor	F1	P	R	Exact Match
CRF	0.620	0.620	0.610	0.245
BERT-base	0.871	0.867	0.868	0.673
+ POS	0.873	0.868	0.869	0.673
+ BIO	0.862	0.824	0.832	0.654
+ POS/BIO	0.856	0.816	0.825	0.643
+ BIO/Viterbi	**0.875**	**0.871**	**0.872**	0.708
+ POS/BIO/Viterbi	0.871	0.868	0.869	**0.710**

Table 1: Self-evaluation results.

Extractor	F1	P	R	Exact Match
CRF	0.701	0.694	0.681	0.328
BERT-base	0.942	0.959	0.962	0.773
+ POS	0.942	0.942	0.942	0.779
+ BIO	0.914	0.925	0.912	0.790
+ POS/BIO	0.913	0.923	0.911	0.785
+ BIO/Viterbi	**0.947**	**0.948**	**0.947**	**0.824**
+ POS/BIO/Viterbi	0.942	0.942	0.941	**0.824**

Table 2: Blind test results.

Table 3 shows the performances of the top-3 teams in the official round. The F1-score, Precision, and Recall of the runner-up models are similar to those of the proposed approach. However, the exact match ratio of our proposed approach outperforms those of other participants, which is 8.7% better than that of the method proposed by the second-place team.

Ranking	F1	P	R	Exact Match
1 - Proposed Approach	0.947	0.948	0.947	0.824
2	0.947	0.947	0.947	0.737
3	0.837	0.836	0.839	0.704

Table 3: Results of official round.

5 Error Analysis

Figure 1 shows an instance for error analysis, where our approach outperforms the baseline models in the self-evaluation experiment. Here we use the last partition of the training data as validation set for example. In this case, there are two causal chains. The CRF model does not find any causal event. The BERT model only succeeds in tagging one of the causal chains, but fails to tag the other effect span "That stake...". In contrast, our approach successfully labels the correct causal events, which shows that it can better identify the proper event span and achieve a higher exact match ratio.

Figure 2 shows an instance that our best model fails to correctly identify the causal events. In this example, our model could not extract the two correct causal events in one sentence, suggesting that our system still has room for improvement on detecting multiple causal events in a single sentence.

: Cause : Effect	**(0237.00009.2)**
Baseline BERT-base	Avid Technology (AVID) Impactive Capital disclosed on Sept. 6 an initial position of 3,665,256 shares of the audio- and video-content maker and distributor. That stake counts 2,528,227 shares purchased at $5.98 to $9.97 each during the period of July 8 through Sept. 6. Impactive said that it acquired the shares because it was an attractive investment.
Our Approach	Avid Technology (AVID) Impactive Capital disclosed on Sept. 6 an initial position of 3,665,256 shares of the audio- and video-content maker and distributor. That stake counts 2,528,227 shares purchased at $5.98 to $9.97 each during the period of July 8 through Sept. 6. Impactive said that it acquired the shares because it was an attractive investment.

Figure 1: Instance for error analysis that our approach succeeds.

: Cause : Effect	**(0151.00014.1) & (0151.00014.2)**
Gold	If the total income of a company was in excess of INR 1 Crore but less than INR 10 Crores then applicable surcharge was of 7% and if income exceeded INR 10 Crores, then the surcharge was charged at 12%.
Our Approach	If the total income of a company was in excess of INR 1 Crore but less than INR 10 Crores then applicable surcharge was of 7% and if income exceeded INR 10 Crores, then the surcharge was charged at 12%.

Figure 2: Instance for error analysis that our approach fails.

6 Conclusion

This paper presents our approach to causality detection, which is ranked first in the Task 2 of FinCausal-2020. The ablation analysis shows the effectiveness of the proposed approach. The error analysis also supports that the proposed approach performs better in multiple causal events and multiple causal chains cases.

In the future, we plan to adopt the proposed approach to the documents in other domains. We also plan to adopt the extracted causality from financial documents to improve the performance of downstream tasks such as stock movement prediction and financial argument mining.

Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 109-2218-E-009-014, MOST 109-2634-F-002-040, and MOST 109-2634-F-002-034, and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

References

Ananth Balashankar, Sunandan Chakraborty, Samuel Fraiberger, and Lakshminarayanan Subramanian. 2019. Identifying predictive causal factors from news streams. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2338–2348, Hong Kong, China, November. Association for Computational Linguistics.

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*.

Kiyoshi Izumi and Hiroki Sakaji. 2019. Economic causal-chain search using text mining technology. In *Proceed-*

ings of the First Workshop on Financial Technology and Natural Language Processing, pages 61–65, Macao, China, August.

Christopher S. G. Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 336–343, Hong Kong, October. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 138–148, Austin, Texas, November. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.