A Generative Approach for Financial Causality Extraction

Tapas Nayak
tnk02.05@gmail.com
IIT Kharagpur
India

Soumya Sharma
soumyasharma20@gmail.com
IIT Kharagpur
India

Yash Butala
butalayash99@gmail.com
IIT Kharagpur
India

Koustuv Dasgupta
Koustuv.x.Dasgupta@gs.com
Goldman Sachs
India

Pawan Goyal
pawang@cse.iitkgp.ac.in
IIT Kharagpur
India

Niloy Ganguly
niloy@cse.iitkgp.ac.in
IIT Kharagpur, India
Leibniz University, Germany

ABSTRACT

Causality represents the foremost relation between events in financial documents such as financial news articles, financial reports. Each financial causality contains a cause span and an effect span. Previous works proposed sequence labeling approaches to solve this task. But sequence labeling models find it difficult to extract multiple causalities and overlapping causalities from the text segments. In this paper, we explore a generative approach for causality extraction using the encoder-decoder framework and pointer networks. We use a causality dataset from the financial domain, *FinCausal*, for our experiments and our proposed framework achieves very competitive performance on this dataset.

CCS CONCEPTS

• **Deep learning** → **Generative models**; • **Natural language processing** → *Information extraction*; • **Event extraction** → causality extraction.

KEYWORDS

financial information extraction, financial causality extraction, generative models, pointer networks

1 INTRODUCTION

Causality extraction from financial text is an important task for the analysis of financial documents such as financial news articles, financial reports. Previously, sequence labeling models [1, 5, 6] were proposed to solve this task and they assign 'BIO' tags for cause and effect span to each token in the text. Although these models perform quite well in this task, they are not modeled to handle the challenges

of multiple and overlapping causalities present in the financial text. In this paper, we explore a generative approach using an encoder-decoder framework for causality extraction. We incorporate pointer networks into our decoding framework for structured prediction of the cause and effects spans in the text. The encoder-decoder approach extracts the causalities in a sequence thus the challenges of variable-length causality extraction and overlapping causalities extraction are solved. The pointer network-based decoding identifies the cause and effect spans using the start and end positional index in the text. So the cause and effect spans of different lengths are modeled uniformly in this approach. We use *FinCausal* [7], a dataset containing text segments from financial news articles, for our experiments. Our proposed model achieves very competitive performance in this dataset. We release our code and data for future research at https://github.com/nayakt/CEPN.

2 THE CEPN FRAMEWORK

We present **CEPN**, a **C**ausality **E**xtraction framework using **P**ointer **N**etwork-based encoder-decoder model. To formally define this task, given a text segment $S = \{w_1, w_2, ..., w_n\}$ with n tokens, the goal is to extract a set of causalities $T = \{y_1, y_2, ..., y_{m-1}\}$ where $y_t = (c_s^t, c_e^t, e_s^t, e_e^t)$ is the tth causality, $m-1$ is the number of causalities. $c_s^t, c_e^t, e_s^t,$ and e_e^t represent the positions of the start and end tokens of the cause and effect span of the tth causality in the text segment S, respectively. In Fig 1, we give an overview of our proposed model which is inspired from the models proposed in similar structure prediction tasks such as joint entity-relation extraction [9] and aspect-sentiment triplet extraction [8].

We use a pre-trained BERT model [3] to encode the source text. We concatenate the part-of-speech tag embeddings of the tokens with the BERT vectors to obtain the encoder hidden states $\mathbf{h}_i^E \in \mathbb{R}^{d_h}$. We add a special token '[unused0]' at the front of the text and use its positional index '0' to stop the decoding process. We use BERT_Base_Cased of dimension 768 and BERT_Large_Cased of dimension $1,024$ model for encoding, and we refer them as CEPN_Base and CEPN_Large. We set the dimension of the part-of-speech tag embeddings at 32, and initialize them randomly. The hidden dimension d_h of encoder for CEPN_Base and CEPN_Large model is $768 + 32 = 800$ and $1,024 + 32 = 1,056$, respectively.

We consider the causalities as a sequence $T = \{y_1, y_2, ..., y_{m-1}\}$. We use two special tuples to model this task as a sequence generation problem. We start the sequence generation process with a special tuple y_0, and we mark the end of this generation process with another special tuple, $y_m = (0, -1, -1, -1)$. Here, only $c_s^t = 0$
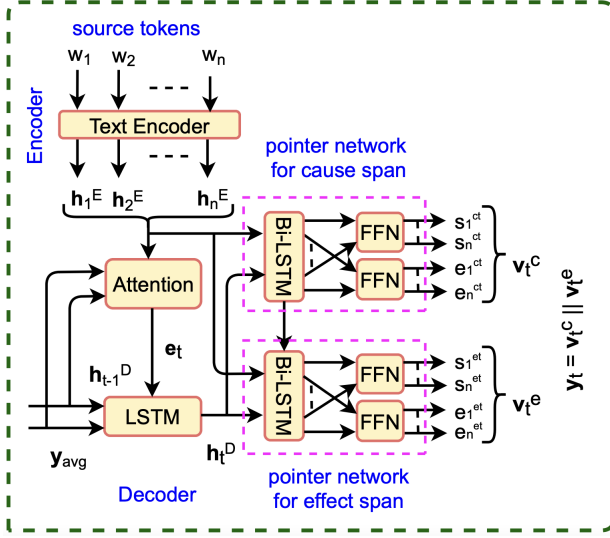
Figure 1: The architecture of CEPN framework.

(corresponds to [unused0] token) is enough to mark the end of the sequence generation process, and we ignore the other indexes of value −1. In the decoder, we use an LSTM cell for generating the causality sequence. We set the hidden dimension of this LSTM cell at d_h that is the same as the hidden dimension of the encoder vectors. We pass the encoder context representation, $e_t \in \mathbb{R}^{d_h}$, and the average of the vectors of already generated causalities, y_{avg} to this LSTM cell to generate the next hidden state, $h_t^D \in \mathbb{R}^{d_h}$. We use two pointer networks for identifying the cause and effect span of the causalities. Each pointer network consists of a BiLSTM layer with hidden dimension d_p and two feed-forward layers (FFN) of dimension $d_p \times 1$ with Softmax activation. We set $d_p = d_h$ as the hidden dimension of the BiLSTM layer in the pointer networks. The output of this FFN/Softmax layer is a scalar score corresponding to each token in the source text. These scalar scores represent the probability distribution of the source tokens being the start or end token of the cause/effect span. We can extract either of the cause or effect spans first, then use the vector representation of that span to extract the other span of a causality. In one version, we extract cause span first, we refer to them using 'CF' such as CEPN_Base_CF. We experiment with another version where we extract the effect span first, and we refer to them using 'EF' such as CEPN_Base_EF.

3 EXPERIMENTS

We use the *FinCausal* 2020[1] and *FinCausal* 2021[2] datasets for our experiments. The text segments in these two datasets contain a maximum of three sentences and the cause and effect span of causality can appear in different sentences in the segment. The number of average tokens in the text segments is ∼ 40 and in the cause and effect span is ∼ 17-18. As the gold-label causalities of the blind test split are not publicly available, we do a 5-fold cross-validation on the training datasets for experiments. We compare our model

[1]http://wp.lancs.ac.uk/cfie/fincausal2020/
[2]http://wp.lancs.ac.uk/cfie/fincausal2021/

against the previous SOTA models such as UPB [4], DOMINO [2], PAMNet [10], GBe [1], NTUNLPL [5]. We report token-level F1 score and exact-match F1 score for comparison.

We include the results on *FinCausal* 2020 and *FinCausal* 2021 datasets in Table 1. We see that our CF and EF versions of the model perform almost equally on both of these datasets. Compare to GBe_Large model, our CEPN_Large_EF model achieves 1.4% higher token-level F1 score and 3.3% higher exact match-based F1 score on the *FinCausal* 2020 dataset. On the *FinCausal* 2021 dataset, compare to GBe_Large model, our CEPN_Large_CF model achieves 0.9% higher token-level F1 score and 1.4% higher exact match-based F1 score in this dataset. We also perform a statistical significance test (two-tailed and paired) between CEPN_Large_CF model and GBe_Large model on the *FinCausal* 2020 and 2021 datasets and find that our model is statistically significant with $p < 0.001$ (achieves a 1.1% higher mean token-level F1 score on both datasets).

	FinCausal 2020		FinCausal 2021	
Model	Token_F1	EM_F1	Token_F1	EM_F1
UPB_Base [4]	0.689	-	-	-
DOMINO_Base [2]	0.796	-	-	-
PAMNet_Large* [10]	0.767	0.485	0.776	0.533
GBe_Large* [1]	0.860	0.710	0.883	0.761
NTUNLPL_Base [5]	0.869	-	-	-
CEPN_Base_CF	0.866	0.733	0.885	0.769
CEPN_Base_EF	0.857	0.722	0.883	0.761
CEPN_Large_CF	0.870	0.739	**0.892**	**0.775**
CEPN_Large_EF	**0.874**	**0.743**	0.891	0.767

Table 1: Performance comparison of CEPN model on the *FinCausal* datasets against the previous SOTA models. * marked baseline scores are reproduced by us. The remaining baseline scores are taken from their papers. 'Base' and 'Large' refer to the use of BERT_Base and BERT_Large models for source encoding. We report the average of the 5-folds in a cross-validation setting.

4 CONCLUSION

In this paper, we explore a generative approach for causality extraction in financial text using the encoder-decoder framework and pointer networks. Our model is designed to handle different challenges in this task such as extracting unknown variable-length causalities, identifying text with no causality, and extracting overlapping causalities from the text in an end-to-end architecture. Experimental results on the *FinCausal* datasets show the effectiveness of our proposed framework on financial causality extraction.

REFERENCES

[1] Guillaume Becquin. 2020. GBe at FinCausal 2020, Task 2: Span-based Causality Extraction for Financial Documents. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation.*

[2] Sharanya Chakravarthy, Tushar Kanakagiri, Karthik Radhakrishnan, and Anjana Umapathy. 2020. Domino at FinCausal 2020, Task 1 and 2: Causal Extraction System. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation.* https://aclanthology.org/2020.fnp-1.15

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL.*

[4] Marius Ionescu, Andrei-Marius Avram, George-Andrei Dima, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. UPB at FinCausal-2020, Tasks 1 & 2: Causality Analysis in Financial Documents using Pretrained Language Models. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation.* https://aclanthology.org/2020.fnp-1.8

[5] Pei-Wei Kao, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. NTUNLPL at FinCausal 2020, Task 2: Improving Causality Detection Using

[6] Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality Extraction based on Self-Attentive BiLSTM-CRF with Transferred Embeddings. *Neurocomputing* (2021).

[7] Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. Financial Document Causality Detection Shared Task (FinCausal 2020). *arXiv preprint arXiv:2012.02505* (2020).

[8] Rajdeep Mukherjee, Tapas Nayak, Yash Butala, Sourangshu Bhattacharya, and Pawan Goyal. 2021. PASTE: A Tagging-Free Decoding Framework Using Pointer Networks for Aspect Sentiment Triplet Extraction. In *EMNLP.*

[9] Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *AAAI.*

[10] Zsolt Szántó and Gábor Berend. 2020. ProsperAMnet at FinCausal 2020, Task 1 & 2: Modeling causality in financial texts using multi-headed transformers. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation.*

Viterbi Decoder. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation.*