LIORI at FinCausal 2020, Tasks 1 & 2

Adis Davletov
RANEPA
Lomonosov Moscow State University
davletov-aa@ranepa.ru

Denis Gordeev
RANEPA
gordeev-di@ranepa.ru

Alexey Rey
RANEPA
rey-ai@ranepa.ru

Nikolay Arefyev
Lomonosov Moscow State University,
Samsung R&D Institute Russia,
National Research University
Higher School of Economics
nick.arefyev@gmail.com

Abstract

In this paper, we describe the results of team LIORI at the FinCausal 2020 Shared task held as a part of the 1st Joint Workshop on Financial Narrative Processing and MultiLingual Financial Summarisation. The shared task consisted of two subtasks: 1) classifying whether a sentence contains any causality and 2) labelling phrases that indicate causes and consequences. We used Transformer-based models with joint-task learning and their voting ensembles. Our team ranked 1st in the first subtask and 4th in the second one.

1 Introduction

The Financial Document Causality Detection Task was devoted to finding causes and consequences in financial news (Mariko et al., 2020). This task is relevant for information retrieval and economics. This task was focused on causality associated with a financial event while an event was "defined as the arising or emergence of a new object or context in regard to a previous situation".

The shared task consisted of two subtasks:

- Sentence Classification

  It was a binary classification task. The goal of this subtask was to detect whether a sentence displayed any causal meanings or not

- Cause and Effect Detection

  This task was a relation detection task. Participants needed to identify "in a causal sentence or text block the causal elements and the consequential ones" [1]. This task could be considered as a sequence labelling problem because individual words and phrases corresponded to three labels: cause, consequence, empty label. Each word or character corresponded to only one label.

For both tasks simultaneously we used a single Transformer-based model (Vaswani et al., 2017) with two inputs and outputs for each of the tasks respectively. The first task was treated as a classification task with a single label for the input, while for the second the label was predicted for each input word. The training and dataset processing code is published on our GitHub page [2].

Our team ranked 1st in the first subtask and 4th in the second one.

[1] http://wp.lancs.ac.uk/cfie/fincausal2020/
[2] https://github.com/InstituteForIndustrialEconomics/fincausal-2020

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 45–49
Barcelona, Spain (Online), December 12, 2020.

2 Related Work

There are many works devoted to sequence labelling in various domains as it is one of the most popular tasks in Natural Language Processing (NLP).

Causality detection in texts is also a very old topic. First works date back to the 80s according to the report by Asghar (Asghar, 2016). Recently there have appeared works that leverage neural networks against for causality labelling (Li et al., 2019). The results of neural networks there seem to be in line with the performance for other sequence labelling tasks such as named entity recognition (Ghaddar and Langlais, 2018) for Bi-LSTM models according to paperswithcode.com [3]. For our work, we adopted a Transformer-based approach as it performs the best against current models for sequence labelling and relation extraction. For example, if we look again at named entity recognition (one of the most popular sequence labelling tasks) - at paperswithcode.com[4], we can see that the top 3 best performing use an attention-based model for Ontonotes v5 and CoNLL 2003. Some recent works have also shown that multi-task learning can produce better results if we have several targets for the same input due to eavesdropping and lower task-bias (Ruder, 2017), thus discouraging model from over-fitting. Recent competitions, where multi-task models perform well, also prove this point (Dai et al., 2020; Davletov et al., 2020; Gordeev and Lykova, 2020).

3 Dataset

The task dataset has been extracted from different 2019 financial news provided by Qwam [5]. The corpus consists of HTML-pages of financial news from 2019. It also contains various financial and legal reports from the SEC Edgar Database ticker list, filtered on financial keywords.

The texts have been normalized for the research task in the following way:

- First, the text was split into sentences.
- Then, sentences containing causal elements were identified.
- The document text is then split into passages of consecutive sentences, keeping causally-related sentences in the same passage which are used for binary predictions in the first subtask.
- Passages with positive classes are used as the dataset for the second subtask.
- The organizers provide the start and end indices for causes and effects.

The dataset was split into trial, train and test datasets by the organizers. The trial and train parts contained training labels, while the test part did not include them and was used for ranking. We combined the trial and train parts and used 20% of the combined dataset for validation.

4 Solution

In this work, we went with multitask Transformer-based models for both subtasks. It means that we had two inputs and outputs, for each of the tasks respectively. In this work we tried BERT (Devlin et al., 2018) and ROBERTa (Liu et al., 2019) based models. BERT is a multilingual language model based on self-attention. ROBERTa is a "robustly optimized" BERT variant with larger mini-batches and byte-level BPE (byte-pair encodings). In both cases we used English large model variants (bert-large and roberta-large). On top of pre-trained BERT and ROBERTa models, we added two Linear layers with dropout for each of the tasks. Cross-entropy was used for training the models. Thus, we had two loss functions (for each of the output layers) that were weighted and concatenated. All used models were provided by

[3] https://paperswithcode.com/sota/named-entity-recognition-ner-on-ontonotes-v5

[4] https://paperswithcode.com/task/named-entity-recognition-ner

[5] http://www.qwamci.com/

Hugging Face (Wolf et al., 2019). Our combined loss function can be seen below, where L_a is the first subtask loss and L_b is the second subtask loss.

$$L_a = -\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{N_c} y_i \cdot log(\hat{y}_i)$$

where m is the number of samples in the batch, y_i is the target value, $\hat{y}_i$ – our predicted value and N_c is the number of classes.

$$L_b = -\frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_j}\sum_{j=1}^{N_j}\sum_{c=1}^{N_c} y_c \cdot log(\hat{y}_c)$$

where m is the number of samples in the batch, N_j is the number of tokens in the batch, N_c is the number of NER classes, $\hat{y}_c$ – the predicted NER class and y_c is the target value.

$\mathcal{L} = \lambda_a L_a + \lambda_b L_b$, where λ are scalar weights for the loss functions.

All padded words and non-labeled words (and their resulting tokens) were excluded from loss function calculation and not included into N_j, while special '[SEP]' and '[CLS]' tokens were included.

While training models for the first subtask we tested a number of weighting schemes ranging between 2 and 0 for sequence labelling subtask loss. However, for the second subtask, the weights for text classification loss were set to zero which makes the model equivalent to a general sequence labelling model. We also tried various sequence labelling formats of the second subtask input: BIO (beginning, inside, outside) and BIEO (beginning, inside, end, outside). Learning rates in the range between $5e-06$ and $5e-05$ were tested. Dropout coefficients were tested from 0.1 to 0.2. For the first subtask, there were also provided the results for ensembles of the best 3, 4 and 5 performing models according to the validation dataset. Simple voting ensembles were used.

We used a system with 2 NVidia RTX2080 GPUs and Google Colab to train all models.

5 Results

Test Score	Validation Score	Model	Target Format	Learning Rate	Text Loss Weight	Sequence Loss Weight	Dropout Rate
0.96529	0.960016	bert	bieo	1e-05	1.0	0.2	0.1
0.965454	0.960291	bert	se	7e-06	1.0	0.1	0.1
0.96685	0.961945	bert	se	5e-05	1.0	0.2	0.15
...	...	...	...	...	...	...	...
0.973839	0.961179	roberta	bio	1e-05	1.0	0.1	0.1
0.973839	0.967221	roberta	bio	1e-05	1.0	0.1	0.1
0.975088	*0.9657*	*roberta*	*bio*	*5e-06*	*1.0*	*0.1*	*0.1*
0.975238		top-3 Ensemble					
0.975735		top-4 Ensemble					
0.977467		**top-5 Ensemble**					

Table 1: Model results for Subtask 1: Sentence Classification. In the Table we provide the results for only the best and the worst 3 models and of the ensembles of the top-N performing models. The results are sorted from the bottom to the top.

For the first subtask, the organizers used F1-score. For the second subtask, the metric is a weighted average F1 score, where the F1 score of each class is balanced by the number of items in each class (see (Mariko et al., 2020)).

In the first subtask our final model achieved F1 equal to 0.977 on the leaderboard (the next participant's score is 0.975 F1), in the second subtask our result was 0.826 F1 with the winning solution having 0.947

Test Score	Validation Score	Model	Target Format	Learning Rate	Text Loss Weight	Sequence Loss Weight	Dropout Rate
0.754986	0.872582	roberta	bio	0.0001	0.0	1.0	0.1
0.76584	0.82897	roberta	bio	0.0001	0.0	1.0	0.2
0.794089	0.865707	roberta	bio	9e-05	0.0	1.0	0.2
...	...	...	...	...	...	...	...
0.823952	0.898873	bert	bio	0.0001	0.0	1.0	0.2
0.824818	0.894067	bert	bio	7e-05	0.0	1.0	0.2
0.826049	**0.906328**	**bert**	**bio**	**0.0001**	**0.0**	**1.0**	**0.1**

Table 2: Model results for Subtask 2: Cause and Effect Detection. In the Table, there are provided the results for only the best and the worst 3 models. The results are sorted from the bottom to the top.

F1. The results of individual models and their hyperparameters can be seen in Tables 1 and 2 for each of the subtasks respectively.

As can be seen from Table 1 for subtask 1 ROBERTa robustly outperforms BERT for the first subtask. The best top-3 single models are ROBERTa-based with various hyperparameters. It can also be seen that sequence loss improves model results, but the best models have their weights scaled down by 0.1. It also should be noted that the difference between all individual models is small and the difference between the best and the worst-performing ones is less than 0.1 F-1-score point. For the first subtask, we also tried an ensemble of 3,4 and 5 best performing individual models. The increase in the number of the used best models consistently improved the results. Thus, it may be also beneficial to train other types of models or to increase the number of models in an ensemble.

Paradoxically, for the second subtask BERT-based models consistently outperform ROBERTa based ones. Moreover, the difference is much larger and constitutes more than 0.7 F1-score points. We did not try ensemble-based models for the second subtask. It also can be seen that all our models tend to overfit to the training and validation datasets. A more robust training scheme such as k-fold cross validation might be of benefit here.

6 Conclusion

This paper describes the results of team LIORI at the FinCausal 2020 Shared task held as a part of the 1st Joint Workshop on Financial Narrative Processing and MultiLingual Financial Summarisation. The shared task consisted of two subtasks: classifying whether a sentence contains any causality and labelling phrases which indicate causes and consequences. Transformer-based models with joint-task learning were used. In this paper we show that different model architectures perform better for different subtasks and that joint-task learning might improve results for some subtasks. However, it also results in slight overfitting for sequence labelling task and might require further investigation.

Acknowledgements

We thank the organisers of the competition for such an inspiring task. We are grateful to our reviewers for their useful suggestions. The contribution of Nikolay Arefyev to the paper was partially done within the framework of the HSE University Basic Research Program funded by the Russian Academic Excellence Project '5-100'.

References

Nabiha Asghar. 2016. Automatic extraction of causal relations from natural language texts: a comprehensive survey. *arXiv preprint arXiv:1605.07895*.

Wenliang Dai, Tiezheng Yu, Zihan Liu, and Pascale Fung. 2020. Kungfupanda at semeval-2020 task 12: Bert-based multi-task learning for offensive language detection. *arXiv preprint arXiv:2004.13432*.

Adis Davletov, Denis Gordeev, Alexey Rey, and Nikolay Arefyev. 2020. Renersans: Relation extraction and named entity recognition as sequence annotation. In *Computational Linguistics and Intellectual Technologies*, pages 187–197.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. oct.

Abbas Ghaddar and Philippe Langlais. 2018. Robust lexical features for improved neural network named-entity recognition. *arXiv preprint arXiv:1806.03489*.

Denis Gordeev and Olga Lykova. 2020. Bert of all trades, master of some. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 93–98.

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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arxiv.org*.

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.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Adv. Neural Inf. Process. Syst.*, volume 2017-Decem, pages 5999–6009.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.0.