Financial Document Causality Detection Shared Task (FinCausal 2020)

Dominique Mariko[1] **Hanna Abi-Akl** [1] **Estelle Labidurie**[1]
Stéphane Durfort[1] **Hugues de Mazancourt**[1] **Mahmoud El-Haj**[2]

[1]YseopLab, FR, [2]Lancaster University, UK,
[1]lab@yseop.com, [2]m.el-haj@lancaster.ac.uk

Abstract

We present the FinCausal 2020 Shared Task on Causality Detection in Financial Documents and the associated FinCausal dataset, and discuss the participating systems and results. Two sub-tasks are proposed: a binary classification task (Task 1) and a relation extraction task (Task 2). A total of 16 teams submitted runs across the two Tasks and 13 of them contributed with a system description paper. This workshop is associated to the Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP-FNS 2020[1]), held at The 28th International Conference on Computational Linguistics (COLING'2020[2]), Barcelona, Spain on September 12, 2020.

1 Introduction

In an effort to automatically interpret the semantics of written languages, the analysis and understanding of causal relationships between facts stand as a key element. A major difficulty regarding automation is that causality can be expressed using many different syntactic patterns as well as contrasted semantic representations. This difficulty is reinforced by the existence of both explicit and implicit cause-effect links. Early works in this field, such as (Khoo et al., 1998), aim at detecting causal relations using linguistic patterns. However, these applications are often restricted to a specific domain, limited to explicit causal relationships only (causal links, causative verbs, resultative constructions, conditionals and causative adverbs and adjectives), and do not take into account the ambiguities of the connectors. The semi-automatic method developed by (Girju and Moldovan, 2002) goes a step further by creating lexico-syntactic patterns based on WordNet semantic relations between nouns, then using semantics constraints to test ambiguous causal relations. Despite better results, the exclusive use of linguistic patterns prevents a fully efficient coverage of all cause-effect links. Consequently, various machine learning techniques were tested for this task. (Chang and Choi, 2004) developed Naive Bayes causality extraction models based on lexical pair probability and cue phrase probability. By focusing on the dynamics of causal relationships, the system PREPOST developed by (Sil et al., 2010) stands as a viable system to detect causal relationships between one event and a consequent state of this event, training a classifier to identify events' preconditions and/or postconditions. In parallel, hybrid methods were also developed such as the expanded semantic parsing of (CMU, 2018). This system combines an SCL approach, pattern-based methods and a neural network architecture, offering more flexibility than exclusive pattern based approaches. Overall, the management of linguistic ambiguities as well as the existence of implicit connections appear to be the main brakes in the identification and extraction of causal relationships.

In this paper, we present the FinCausal Corpus and the associated featured Tasks, as a contribution to the research effort addressing implicit and multiple causalities detection automation in financial documents. All the datasets created for this shared task are publicly available to support further research on Causality modelling[3], and the detailed annotation scheme is provided in the Appendix A. Next, Sec-

[1]http://wp.lancs.ac.uk/cfie/fnp2020/
[2]https://coling2020.org
[3]https://competitions.codalab.org/competitions/25340

tion 2 describes the FinCausal Corpus and Section 3 presents the Tasks. Section 4 provides the baseline proposed to participants and details their results, with a high-level description of the approaches they adopted. Finally, Section 5 concludes this report and discusses some future directions.

2 FinCausal Corpus

The data are extracted from a corpus of 2019 financial news provided by Qwam, collected on 14.000 economics and finance websites. The original raw corpus is an ensemble of HTML pages corresponding to daily information retrieval from financial news feed. These news mostly inform on the 2019 financial landscape, but can also contain information related to politics, micro economics or other topic considered relevant for finance information. Data are released under the CC0 License[4].

All collected HTML files were initially split into sentences according to their punctuation, then were grouped into text sections of 1 to 3 sentences after the data annotation process has been completed. Below are the principle global metrics gathered during the annotation process. The metrics are defined with respect to the annotation scheme.

For consistency in our references, we refer to a **file** as the original document to process, a **text section** as a multi-sentenced text string (1 to 3 sentences that may or may not contain causality) and a **chunk** as a substring (consisting either of a part of a sentence, a whole sentence or multiple sentences) within a text section. We also retain statistics related to the main tags used in our annotation scheme during the preparation of the datasets. These tags are defined as follows:

- **Cause**: Indicates the presence of causality
- **QFact**: Qualifies the causal chunk as quantitative (i.e., containing numerical entities like amounts)
- **Fact**: Qualifies the causal chunk as non-quantitative
- **Discard/Remove**: Indicates text that is not retained for the final datasets (non financial texts)

In addition, metrics related to **fact alignment** (i.e., trimming sentences according to preset priority rules in the annotation scheme) are also included to consistently reflect the preprocessing carried out at this step. All statistics are provided for the 3 datasets provided to participants: Trial, Practice, Evaluation as well as global statistics to present a general outlook on the overall annotation phase. The resulting statistics are collected in Table 1.

Metric	**Trial**	**Practice**	**Evaluation**	**Global**
Total annotated files	695	832	1878	3405
Total sentences in files before definition of text sections	25326	29381	74951	129658
Total **Cause** tags in files	657	1128	2244	4029
Total **QFact** tags in files	937	1824	2589	5350
Total **Fact** tags in files	449	999	2514	3962
Total **Discard/Remove** tags in files	1030	612	2462	4104
Total files in review for **fact alignment**	375	560	705	1640
Total files modified in **fact alignment**	116	182	259	557
Average causalities per file	2.73	3.06	2.98	2.92
Average offset of 2nd sentence in text sections	137	139	141	139
Average offset of 3rd sentence in text sections	270	277	282	276
Percentage of multi-sentenced text sections	59.23	51.02	37.52	49.26

Table 1: Global Distribution of Annotated files

After fact alignment and inter annotator agreement (see Appendix A), a Trial and Training sets with Gold annotations were released, along with a blind Evaluation set for systems evaluation.

[4] https://creativecommons.org/publicdomain/zero/1.0/deed.en

3 Tasks

Both substaks are intented as a pipeline. The first one aims at detecting if a text section contains a causal scheme (as defined in Appendix A.1), the second one aims at identifying cause and effect in a causal text section. Participants were allowed to concatenate and split the Trial and Practice datasets as they saw fit to train their system.

3.1 Task1

Task 1 is a binary classification task. The dataset consists of a sample of text sections labeled with 1 if the text section is considered containing a causal relation, 0 otherwise. The dataset is by nature unbalanced, as to reflect the proportion of causal sentences extracted from the original news, following the distribution displayed in Table 2.

Metric	**Trial**	**Practice**	**Evaluation**	**Global**
Total number of text sections	8580	13478	7386	29444
Total number of causal text sections	569	1010	567	2136
Percentage of causal text sections	6.63	7.49	7.68	7.24

Table 2: Task 1 Distribution

The trial and practice samples were provided to participants as csv files with headers *Index; Text; Gold*.

- Index: ID of the text section. Is a concatenation of [file increment . text section index]
- Text: Text section extracted from a 2019 news article
- Gold: Gold Label provided from manual annotation

Index	Text	Gold
23.00005	Electric vehicle manufacturers, components for the vehicles, batteries and producers for charging infrastructure who invest over Rs 50 crore and create at least 50 jobs stand eligible for total SGST (State GST) refund on their sales till end of calendar year 2030.	0
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.	1

Table 3: Two examples from FinCausal Task 1 Corpus - Practice dataset

3.2 Task2

The purpose of this task is to extract, in provided text sections, the chunks identifying the causal sequences and the chunks describing the effects. The text sections correspond to the ones labeled as 1 in the Task 1 datasets, except in the blind Evaluation set.

The trial and practice samples were provided to participants as csv files with headers: *Index; Text; Cause; Effect*

- Index: ID of the text section. Is a concatenation of [file increment . text section index]
- Text: Text section extracted from a 2019 news article
- Cause: Chunk referencing the cause of an event (event or related object included)
- Effect: Chunk referencing the effect of the cause

Average statistics on the causes and effects chunks detected in the causal text sections are provided in Table 4. As explained in section 3, complex causal chains are considered during the annotation process, leading to one text section possibly containing multiple causes or effects.

Metric	Trial	Practice	Evaluation	Global Average
Average character length of causal chunks	113.73	109.13	112.48	111.78
Average character length of effect chunks	107.79	104.78	99.66	104.08
Total number of text sections	641	1109	638	796
Total number of unicausal text sections	500	913	452	621.67
Total number of multicausal text sections	141	196	186	174.33

Table 4: Task 2 Distribution

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.	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 5: Three examples from FinCausal Task 2 Corpus - Practice dataset

4 Evaluation

A baseline was provided on the trial samples for both Tasks 1 and 2[5]. Participating systems were ranked on blind Evaluation datasets based on a weighted F1 score, recall, precision for Task 1, plus an additional Exact Match for Task 2. Regarding official ranking, weighted metrics from the scikit-learn package[6] were used for both Tasks, and the official evaluation script is available on Github[7]. Participating teams were allowed to submit as many runs as they wished, while only their highest score was withheld to represent them during evaluation. In addition, they were proposed to enhance their system in a post-evaluation phase[8]. Only the scores validated during the evaluation phase of the competition are displayed below. Amongst the 13 participating teams, six choose to address Tasks 1 and 2 and one (ProsperAMnet) proposed an integrated pipeline for both. Details on the methods and features used by different systems are provided in Table 8 for both Tasks. Noticeably, 7 teams plan to release the code associated to their system publicly.

4.1 Task1

Results for participating teams are provided in Table 6. Last line displays the baseline that had been provided for the task. The baseline was computed using the BERT-base-uncased language model[9] and fine tuned on the Task data using the Hugging Face transformers library (Wolf et al. ., 2019)[10], on a GeForce GTX 1070 8Gb RAM GPU. For Task 1, 6 participants out of 10 took advantage of large Transformers architectures (Vaswani et al. ., 2017) and fine-tuned their systems using the same library as the baseline. Four used Ensemble strategies to aggregate their results and enhance the robustness of their model. Additional strategies such as Data Augmentation and Oversampling are also proposed to work around the unbalanced nature of the data. The best result in terms of weighted-averaged F1-score

[5] https://github.com/yseop/YseopLab/tree/develop/FNP_2020_FinCausal/baseline

[6] https://scikit-learn.org/stable/modules/model_evaluation.html#multiclass-and-multilabel-classification

[7] https://github.com/yseop/YseopLab/tree/develop/FNP_2020_FinCausal/scoring

[8] https://competitions.codalab.org/competitions/25340

[9] https://huggingface.co/bert-base-multilingual-uncased

[10] https://huggingface.co/transformers/model_doc/bert.html

is achieved by the winning team LIORI (97.75%), closely followed by UPB and ProsperAMNet with F1 scores of 97.55% and 97.23%, respectively. The top five systems all leveraged Transformers architectures with associated language models features, evaluating at least on a fine-tuned BERT-base model and providing a comparison with similar models (BERT-large, RoBERTa, and specialized BERT such as FinBERT). The top 2 systems used Ensemble methods (See Table 8). BERT-like systems weighted-F1 ratings are in range [97.75 , 95.78], whereas systems using more traditional Machine Learning models have scores in range [95.00 , 93.09], including systems using BERT-like embeddings in their processing.

Team	F1 Score	Recall	Precision
LIORI	97.75 (1)	97.77 (1)	97.73 (1)
UPB	97.55 (2)	97.59 (2)	97.53 (2)
ProsperAMnet	97.23 (3)	97.20 (3)	97.28 (3)
FiNLP	96.99 (4)	97.03 (4)	96.96 (4)
DOMINO	96.12 (5)	96.06 (5)	96.19 (5)
IIT_kgp	95.78 (6)	95.83 (6)	95.74 (6)
LangResearchLab_NC	95.00 (7)	94.92 (7)	95.08 (7)
NITK NLP	94.35 (8)	94.87 (8)	94.32 (8)
fraunhofer_iais	94.29 (9)	94.76 (9)	94.20 (9)
ISIKUN	93.09 (10)	94.33 (10)	93.89 (10)
baseline	95.23	95.21	95.26

Table 6: Task 1 Results

4.2 Task2

Results for Task 2 are provided in Table 7. Last line displays the baseline that has been provided for Task 2, computed with a CRF model using the pycrfsuite package[11]. One of the challenge of this task was to rebuilt the correct span of causal chunks, according to the annotation scheme. The baseline has been kept deliberately low as is does not take this specific problem into account, nor does it focuses on parameter-tuning strategies, though tuning examples are proposed with the code baseline. All participants decided for sequence labelling strategies and used specific penalization methods and/or heuristics to work around the chunks reconstitution problem. The best performer in this subtask (NTUNLP) uses a BERT-CRF system and a Viterbi decoder for span optimization, achieving (94.72%) weighted F1, closely followed by a BERT-SQUAD augmented system with heuristics for span achieving 94.66% F1 (Gbe).

Team	F1 Score	Recall	Precision	Exact match
NTUNLPL	94.72 (1)	94.70 (1)	94.79 (1)	82.45 (1)
GBe	94.66 (2)	94.66 (2)	94.67 (2)	73.67 (2)
ProsperAMnet	83.71 (3)	83.63 (3)	83.92 (3)	70.38 (4)
LIORI	82.60 (4)	82.80 (4)	82.48 (4)	70.53 (3)
DOMINO	79.60 (5)	78.90 (5)	81.90 (5)	00.00 (7)
fraunhofer_iais	76.00 (6)	74.89 (7)	79.95 (6)	19.12 (5)
JDD	75.61 (7)	75.57 (6)	75.95 (7)	00.00 (7)
UPB	73.10 (8)	72.14 (8)	75.61 (8)	18.34 (6)
baseline	51.06	51.74	50.99	11.11

Table 7: Task 2 Results

[11] https://python-crfsuite.readthedocs.io/en/latest/

Team	F1	Techniques								
		ML	Neural	TF	Ens	AGM	RS	LM	WCS	HS
Task 1										
LIORI	97.75			X	X			X		
UPB	97.55			X	X			X		
ProsperAMnet	97.23			X				X		
FiNLP	96.99			X	X	X	X	X		
DOMINO	96.12			X				X		
IIT_kgp	95.78			X				X		
LangResearchLab_NC	95.00		X				X	X	X	
NITK NLP	94.35	X							X	
fraunhofer_iais	94.29	X			X				X	
ISIKUN	93.09	X							X	
Task 2										
NTUNLPL	94.72	X		X				X	X	
GBe	94.66			X				X		X
ProsperAMnet	83.71			X				X		
LIORI	82.60			X				X		
DOMINO	79.60			X				X		X
fraunhofer_iais	76.00	X	X					X		X
JDD	75.61	X	X					X	X	
UPB	73.10	X						X		X

Table 8: Approaches adopted by the participating teams in Tasks 1 and 2. ML refers to any non-neural machine learning technique such as XGBoost, SVM, etc. Neural refers to any neural network architecture such as BILSTM, CNN, GRUs,etc, except Transformers. TF refers to Transformers architecture. Ens corresponds to Ensemble Learning method. RS is resampling method. HS implies some heuristics has been used in the final computation, mostly to adapt the span in Task 2. LM refers to any language model embedding features. WCS refers to Word, Character or Syntax based features.

5 Conclusion

In this paper, we present the framework and the results for the FinCausal Shared Task. In addition , we present the new FinCausal dataset built specifically for this shared task. We plan to run similar shared tasks in the near future, possibly with some augmented data, in association with the FNP workshop.

Acknowledgements

We would like to thank our dedicated annotators who contributed to the building the FinCausal Corpus: Yagmur Ozturk, Minh Anh Nguyen, Aurélie Nomblot and Lilia Ait Ouarab, as well as the FNP Committee for their gracious support.

References

Avirup Sil, Fei Huang and Alexander Yates. 2010. *Extracting action and event semantics from web text*. AAAI Fall Symposium: Commonsense Knowledge.

Beth Levin and Malka R. Hovav. 1994. *A Preliminary Analysis of Causative Verbs in English*. Lingua 92, 35-77.

Christopher S.G. Khoo, Jaklin Kornfilt, Sung Hyon Myaeng and Robert N. Oddy. 1998. *Automatic Extraction of cause-effect information from newspaper text without knowledge-based inferencing*. Literary & Linguistic Computing 13, 177-186.

Du-Seong Chang and Key-Sun Choi. 2004. *Causal Relation Extraction Using Cue Phrase and Lexical Pair Probabilities*. Natural Language Processing–IJCNLP, 61–70.

Erika Nazaruka. 2019. *Identification of Causal Dependencies by using Natural Language Processing: A Survey*. ENASE 2019.

Jesse Dunietz. 2018. *Annotating and Automatically Tagging Constructions of Causal Language*. Carnegie Mellon University.

Jesse Dunietz, Lori Levin, Jaime Carbonell. 2017. *The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations*. Proceedings of the 11th Linguistic Annotation Workshop, ACL Anthology 2017.

Nabiha Ashgar. 2016. *Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey*. Arxiv 2016.

Roxana Girju and Dan Moldovan. 2002. *Text mining for causal relations*. FLAIRS Conference, 360–364.

Ashish Vaswani, Noam Shazeer, Niki Parmar et al. 2017. *Attention is All you Need*. Advances in Neural Information Processing Systems 30, 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh et al. 2019. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. Arxiv, abs/1910.03771.

Appendix A. Annotation scheme

In this appendix, we provide detailed information on the concepts guiding the annotation. The annotation process was iterative: Annotations were proposed on a BRAT annotation server[12] by a first annotator then revised by two others until agreement. These agreement sessions were the opportunity to define and iterate on the following annotation scheme.

A.1. Defining causatives

A causal relationship involves the statement of a **cause** and its **effect**, meaning that two events or actors are related to each other with one triggering the other. We focused our annotation on text sections[13] that state causal relationships involving a quantified fact, which was necessary to reduce the complexity of the task. Table 9 displays the terms used in the context of the Shared Task.

FACT	
Empirical Fact	Past event, acknowledged
Process	Concrete event in duration
State of Affairs	In being situation (will become true or false)
Looking Forward Statement	Expectation (often a declaration made by CPY board member)
Hypothesis	Projection based on facts
QUANTIFIED FACT (QFact)	
Explicit	Has a direct connection to an explicit measure
Measurable	Measure is either a quantity or a number that can be precisely identified in a text section
Verifiable	Is a State of Affairs at least

Table 9: Representation of events terminology

In this scheme, an effect can only be a quantified fact. The cause can either be a fact or a quantified fact. The causality between these two elements can be implicit as well as explicitly stated with a triggering linguistic mark also called a connective. The place of these chunks in the text section can vary according to the connective used or simply according to the author's style.

In order to delimit the process, the distance between a cause fact and an effect fact was restricted to **a 3-sentences distance**. In other words, we only annotated a causal relationship when there was a maximal gap of 1 untagged sentence between the two facts. For instance, in the text section *<cause>Previous*

[12] https://brat.nlplab.org/installation.html

[13] We are using the term *text section* since it could be a phrase, a sentence as well as a paragraph in which the cause and the effect are split in different sentences. For instance "Selling and marketing expenses decreased to $1,500,000 in 2010. This was primarily attributable to employee-related actions and lower travel costs". However, in order to have a reproducible annotation process, we reduced the context to a paragraph of maximum three sentences.

management sought to transform the company from a simple milk processor into a producer of value-added dairy products as it chased profits offshore<cause>.<effect>Among Fonterra's biggest missteps was the 2015 purchase of an 18.8 per cent stake in Chinese infant formula manufacturer Beingmate Baby & Child Food for $NZ755 million, just as the China market became hyper-competitive and demand slowed <effect>. Fonterra last month announced it would cut its Beingmate stake by selling shares after failing to find a buyer. Meanwhile, back home, Fonterra's share of the milk processing market dropped from 96 per cent in 2001 to 82 per cent currently, with consultants TDB Advisory expecting it to be about 75 per cent by 2021. In this example, "the 2015 purchase of an 18.8 per cent stake in Chinese infant formula manufacturer Beingmate Baby & Child Food for $NZ755 million" was annotated because the cause and the effect have a 2-sentences distance. On the other hand, "Fonterra's share of the milk processing market dropped from 96 per cent in 2001 to 82 per cent currently" was not annotated because this effect is at a 4-sentences distance from the cause.

A.2. Connectives

A connective can be a verb, a preposition, a conjunction, an element of punctuation, or anything else, which explicitly introduces a causal relationship. Among those, there is a specific type of connective that is not taken into account in this Shared Task called lexical causative (Levin and Hovav, 1994) . A lexical causative is a causal relationship stated through connectives (generally predicates) which, from a semantic point of view, also bear the effect of the cause. We will not consider those as causal references, since the effects are *implied* in the connectives' definition. For instance in "The company raised its provisions by 5% in 2018.", *raise* is a lexical causative that can be glossed as *The company caused the provisions to rise by 5%*.

Causal relationships can be introduced by other types of connectives in the identified text section. It is often rendered with the use of polysemous connectives which main function is not to introduce a causal relationship. For example, in this sentence: "Zhao found himself 60 million yuan indebted after losing 9,000 BTC in a single day (February 10, 2014)", the main function of the connective *after* is to express a temporal relation between the two clauses. But we also have a causal relationship between them, since one triggers the other.

In the tagging process, the connectives involved in the causal relationship **were not annotated as part of the facts**. For example: <effect>*Titan has acquired all of Core Gold's secured debt for $US2.5 million*<effect>in order to <cause>*ensure the long-term success of its assets.*<cause>. The only exception would be when the connective is inserted in the fact. In that case, the connective was annotated. For instance: <cause>*On August 30, 2013, ST Yushun, in order to strengthen its competitive strength*<cause>, <effect>*acquired a 100% stake in ATV Technologies for 154 billion yuan*<effect>.

A.3. Complex causal relationships

In a text section, complex causal relationships can be rendered with conjoined relationships. A conjoined causal relationship can be one cause related to several effects, or one effect caused by several causes. This is often the case when the facts are not repeated and a conjunction is used as a link for the different effects or causes. This phenomenon can be also found in an implicit causal relationship and/or at sentence level. Here is an instance of a conjoined effect related to two causes: <cause>*India's government slashed corporate taxes on Friday* <cause>, <effect>*giving a surprise $20.5 billion break*<effect><cause>*aimed at reviving private investment and lifting growth from a six-year low that has caused job losses and fueled discontent in the countryside*<cause>. In the tagging process, they were all annotated as separate facts apart if a priority rule was to be taken into account.

A.4. Priority rules

The priority rules allow the annotation process of causal relationships to be more accurate and harmonious.

First rule. If a sentence contained **only one fact** (cause or effect), we **tagged the entire sentence** (even if it contains some noise or a connective). For instance: <cause>*Hurricane Irma was the*

most powerful storm ever recorded in the Atlantic and one of the most powerful to hit land, Bonasia said.<cause><effect>It cause $50 billion in damages.<cause>

Second rule. The **annotation of sentence-to-sentence causal relationships is prioritized**. When the annotator had the choice between linking two full sentences together or subdividing a sentence, he chose the sentence-to-sentence annotation. To illustrate this point, let's look at the text section: "*Finally, Seizert Capital Partners LLC increased its holdings in shares of BlackRock Enhanced Global Dividend Trust by 17.2% during the second quarter. Seizert Capital Partners LLC now owns 138,020 shares of the financial services provider's stock valued at $1,481,000 after acquiring an additional 20,223 shares in the last quarter.* In this text section, there are two causal relationships. The first one links "*Seizert Capital Partners LLC increased its holdings in shares of BlackRock Enhanced Global Dividend Trust by 17.2% during the second quarter*" and "*Seizert Capital Partners LLC now owns 138,020 shares of the financial services provider's stock valued at $1,481,000*". Since the two facts are located into different sentences, we would have to annotate the full sentences each time (rule 1). The second causal relationship links "*Seizert Capital Partners LLC now owns 138,020 shares of the financial services provider's stock valued at $1,481,000*" and "*acquiring an additional 20,223 shares in the last quarter*". Here, a sentence is subdivided.

Considering the priority of sentence-to-sentence annotation, the final annotation of this text section was: "<cause>*Finally, Seizert Capital Partners LLC increased its holdings in shares of BlackRock Enhanced Global Dividend Trust by 17.2% during the second quarter*<cause>. <effect>*Seizert Capital Partners LLC now owns 138,020 shares of the financial services provider's stock valued at $1,481,000 after acquiring an additional 20,223 shares in the last quarter.*<effect>".

This rule also highlights the fact that **two different annotations cannot overlay**. It is impossible to annotate "*acquiring an additional 20,223 shares in the last quarter*" and "*Seizert Capital Partners LLC now owns 138,020 shares of the financial services provider's stock valued at $1,481,000 after acquiring an additional 20,223 shares in the last quarter.*" because the same text segment would be part of two different annotations.

Third rule. If a sentence contained **both a cause and an effect**, the **sentence was subdivided**. The spanning was realized so that the exact segments corresponding to the cause and the consequence were selected. For instance: *This week's bad news comes from Rothbury, Michigan, where* <cause>*Barber Steel Foundry will close at the end of the year* <cause>, <effect>*leaving 61 people unemployed*<effect>. However, in the dataset, the spans were extended in order to cover the entirety of the sentence. Only the connector, when located in between the cause and the effect, was left out of the extraction. As a result, in the final dataset we have: <cause>*This week's bad news comes from Rothbury, Michigan, where Barber Steel Foundry will close at the end of the year* <cause>, <effect>*leaving 61 people unemployed*<effect>. The spanning extension facilitate the consistency of the annotation process.

Fourth rule. If **two facts of the same type were located in the same sentence and were related to the same effect or cause**, then **we annotated these two facts as one unit**. For instance, in the text section "*Thomas Cook's demise leaves its German operations hanging. More than 140,000 German holidaymakers have been impacted and tens of thousands of future travel bookings may not be honored.*", the cause fact is "*Thomas Cook's demise*". Since it was the only fact in the sentence, we annotated the full sentence as the cause (see priority rule number 1). The cause fact has two consequences: "*More than 140,000 German holidaymakers have been impacted*" and "*tens of thousands of future travel bookings may not be honored*". Since both effect facts are in the same sentence and related to the same cause, we annotated the text section as follow: <cause>Thomas Cook's demise leaves its German operations hanging.<cause><effect>More than 140,000 German holidaymakers have been impacted and tens of thousands of future travel bookings may not be honored<effect>.

This rule was also applied to the annotation of cause.s and effect.s inside a sentence. For instance:

”<effect>*Our total revenue decreased to $31 million*<effect>due to <cause>*decrease in orders from approximately $91,000 to $82,000, and a decrease in total buyers, which includes both new and repeat buyers from approximately 62,000 to 56,000.*<cause>”. The two causes were put together since they are related to the same effect.

This rule was only used in the two cases presented above. When more than two sentences were involved it was not taken into account. For example: ”<cause>*Let’s say Shirley reduced her assets of $165,000 through a gift of $10,000 and pre-paying her funeral expenses for $15,000.*<cause><effect1>*Her DAC would reduce from $55 a day to $43 a day (a saving of just over $4,300 a year).*<effect1><effect2>*Her equivalent lump sum would reduce by almost $88,000!*<effect2>”. Consequently, the same text section may appear twice in the release dataset.

Fifth rule. The annotation of **causal chains** inside a sentence. A segment of text that is a cause can also be the effect of another cause. For instance, the sentence ”*BHP emitted 14.7m tonnes of carbon dioxide equivalent emissions in its 2019 fiscal year, down from 16.5m tonnes the previous year due to greater use of renewable energy in Chile.*” contains three facts. ”*greater use of renewable energy in Chile* is the cause of *down from 16.5m tonnes the previous year* which is also the cause of *BHP emitted 14.7m tonnes of carbon dioxide equivalent emissions in its 2019 fiscal year*.

In that case, we **isolated the rightmost fact and tagged it according to its nature**. All **the remaining facts were gathered as one unit and annotated with the remaining tag**. In our example it gave the final annotation: ”<effect>*BHP emitted 14.7m tonnes of carbon dioxide equivalent emissions in its 2019 fiscal year, down from 16.5m tonnes the previous year*<effect><cause>*greater use of renewable energy in Chile*<cause>”

A.5. Other annotation levels

The cause or the effect can sometimes be found as pronouns, relative pronouns included. In that case, the reference (the antecedent) of the pronoun, is the extracted element. For instance, in the text: ”The tax revenues decreased by 0.3%, which was caused by fiscal decentralization reform.” *The tax revenue decreased by 0.3%* corresponds to the effect and *fiscal decentralization reform* is the cause. In some cases, the pronoun can be added to the opposite fact where the antecedent is.

The role of a clause in a causal sentence can be ambiguous to identify. For example, it can be precarious to tell whether the clause corresponds to the cause, the means or the goal. If so, the sequence was annotated as the cause.

The ambiguity can also exist between two facts - which is the cause? which is the effect? In that case, when there was only one Qfact, the latter was annotated as the effect. When both facts were Qfacts, the annotation order was left to the annotator’s appreciation. The annotator was encouraged to use reformulation in order to decide which fact was the cause and which fact was the effect.

If the cause is in the middle of the effect or vice versa, the sentence is not annotated because of the conflict process. Here is an example: ”The take-home pay after necessary deductions is S$4,137.” where *after necessary deductions* is a cause inserted in the effect.

We decided not to annotated causal relationships with structures identical to a calculation structure. For instance, in the text section ”*Google has 100K+ people and $136B in revenue (2018), earning over $1.3M per person.*”, we considered that, since the quantified data in the effect fact is the result of a calculation based on the data present in the cause fact, there was no new information. Consequently, there was no need to annotate.

Finally, dates are also to be included in the fact annotated if it is related to it and is placed next to it in the sentence.