Data Processing and Annotation Schemes for FinCausal Shared Task

Yseop Lab
D. Mariko, E. Labidurie,
Y. Ozturk, H. Abi-Akl,
H. de Mazancourt, S. Durfort,
fin.causal.task@gmail.com

Abstract

This document explains the annotation schemes used to label the data for the FinCausal Shared Task. This task is associated to the Joint Workshop on Financial Narrative Processing and Multi-Ling Financial Summarisation (FNP-FNS 2020), to be held at The 28th International Conference on Computational Linguistics (COLING'2020), Barcelona, Spain on the 13 September 2020.

1 Introduction

Causality detection is a well known topic in the NLP and linguistic communities and has many applications in information retrieval. It has been studied extensively in a wide range of disciplines and domains knowledge, yet experts often disagree on the characterisation of a sufficient causal link, as causality can be expressed using many different syntactic patterns as well as contrasted semantic representations. This shared task proposes data to experiment causality detection, and focuses on determining causality associated to an event. An event is defined as the arising or emergence of a new object or context in regard of a previous situation. So the task will emphasise the detection of causality associated with financial or economic analysis and resulting in a quantified output.

Contributors are free to decide on their preferred methodology, would it be regex parsing, corpus linguistics, entity-relationship models, ontology matching, machine or deep learning models, or hybrid methods.

2 Data

2.1 Data Processing

The data are extracted from a corpus of 2019 financial news crawled by Qwam. 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. This raw set has been normalised as to fit in the following format: *Index; Text*

- Index: the ID of the text section.
- Text: Text section extracted from the news. A text section is a block of text from 1 to 3 sentences. The text sections are originated from the same document and are extracted in the following way: the sentence(s) containing causal elements are first identified in the document. The document text is then split into passages of consecutive sentences, keeping causally-related sentences in the same passage.

2.2 Data Release - Provisional Dates

1. Trial data set released on the 1st of February 2020
2. Training data released on the 1st of March 2020

3. The final scores will be evaluated on a blind test dataset, released on the 1st of April 2020

4. Contributions from participants are expected on the 20th of April 2020

5. Release of results are provided by organisers on the 1st of May 2020

2.3 Data License

Data are released under the CC0 License

3 Annotation scheme

3.1 Definition of causality

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[1] that state causal relationships involving a quantified fact, which was necessary to reduce the complexity of the task. The following figure displays the terms we use in the context of the Shared Task.

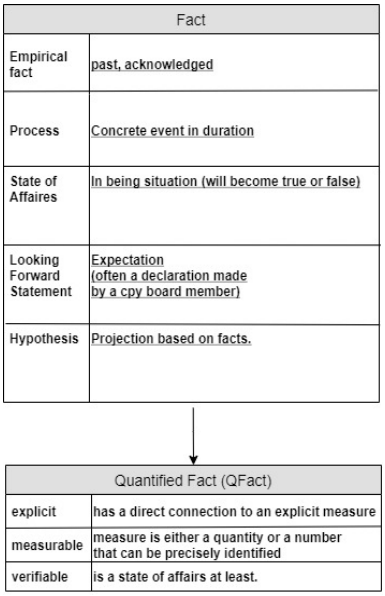

Figure 1: Representation of events terminology

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 sub-strings 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.
As a matter of example:
"<cause>*Previous 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*

[1]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.

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 as effect because this effect is within a 2 sentences away distance from the cause sentence. 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 4 sentences away from the cause.

3.2 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. 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%*. We will not consider those as causal references, since the effects are *implied* in the connectives' definition.

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>. Two exceptions to this scheme are inserting the connectives in the cause or effect:

- the connective is inserted in the annotated fact, i.e. : <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>.

- the connective is at the beginning of a spanned linguistic unit, i.e. <cause>***As*** *Nvidia Com (NVDA) Stock Value Declined*<cause>, <effect>*Shareholder Assetmark Has Cut Stake by $15.67*<effect> (see Third rule in 3.4 Priority Rules)

3.3 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 first annotated as separate facts and then grouped according to priority rules if any applied (see Fourth rule and Fifth rule in 3.4 Priority Rules)

3.4 Priority rules

The priority rules allow the annotation process of causal relationships to be more accurate and harmonious [2].

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.*<effect>

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 that **two different annotations cannot overlap**. We choose not to annotate the following causal chain: "<cause>*acquiring an additional 20,223 shares in the last quarter*"<cause> and <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>, because these two texts segments reconcile into a higher level <effect> in our annotation scheme

Third rule. **When a causal chain is located inside a single sentence, in order to facilitate the extraction process, we chose to span the causal units as much as possible**, i.e. considering the following exact causal units:
"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>",
The spans were extended in order to cover the entire sentence. Only the connector, when located in between the cause and the effect, was left out of the extraction. As a result, the final causal chain is:
<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

[2] As it might sometimes be difficult to distinguish sentences form strings in the designed blocks, we added the index of the offset we used in the Task 2 dataset to segment text blocks into sentences and apply priority rules.

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> <cause>*due to 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 the above example this rule eventually provides this 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>"

3.5 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 fact *earning over $1.3M per person.* is the result of a calculation that can be recomputed from data available in the cause fact, it triggers no new information. Consequently, it was not considered a causal chain and was not annotated.

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.

4 Task 1

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 and SEC corpus, with provisional distribution between 6 and 7 % 1 and 94 and 93% 0.

- The trial sample is a set of 8500 text samples, with same distribution.
- The training sample is a set of 13500 text samples, unbalanced under the same distribution. The trial and training samples will be provided to participants as csv files with headers *Index; Text; Gold.*
  - Index: the ID of the passage
  - Text: Passage extracted from a news article
  - Gold: Gold Label provided from manual annotation

Index	Text	Gold
1	As customer expectations continuously evolve, customers expect immediacy and simplicity.	0
2	Thomas Cook's subsidiary in Germany is still technically operating as of Monday afternoon but has stopped taking bookings. More than 140,000 German holidaymakers have been impacted and tens of thousands of future travel bookings may not be honored	1

Table 1: Task 1 - Training Data Example

- The blind test sample will display the same distribution as trial and training sets.

4.1 Task 1 - Evaluation Metrics

A baseline and an evaluation script for this task will be released with the blind test data. The metrics for final evaluation on the blind test will be precision, recall and a standard F1 score.
We retained the definition of the F1 score as F1 = 2 * (precision * recall)/ (precision + recall)

4.2 Task 1 - Expected Submission

The participants are expected to submit their results by augmenting the blind test dataset with a column containing their predicted results, in a csv file. Their contribution will be scored from the precision, accuracy and F1 metrics provided in 4.1 Task 1 - Evaluation metrics section.

- The blind test data will be provided in a csv file with headers: *Index; Text*
- The expected results would be provided by the participants in a csv file with headers: *Index; Text;* ***Prediction***

605 ; The board of directors of MFA Financial Inc (NYSE:MFA) has approved a regular cash dividend of USD0.20 per share of common stock for the third quarter of 2019, the company disclosed on Thursday. All stockholders of record as on 30 September 2019 will be paid the dividend on 31 October 2019 ; ***1***

627 ; A real estate investment trust, MFA Financial is primarily engaged in the business of investing in residential mortgage assets on a leveraged basis. Its principal business objective is to generate net income for distribution to stockholders. The company was incorporated in Maryland on 24 July 1997 and began operations in 1998. ; ***0***

5 Task 2

Task 2 is a relation extraction task. The text sections will correspond to the ones labeled as 1 in the Task 1 dataset, though for the purpose of results evaluation, they will not be exactly the same in the blind test set. The purpose of this task is to extract, in a causal text section, the sub-string identifying the causal elements and the sub-string describing the effects.

- The trial sample is around 500 samples.
- The training sample is a set of 1200 samples. The trial and training samples will be provided to participants as csv files with headers: *Index; Text; Cause; Effect; Offset_Sentence2, Offset_Sentence3; Cause_Start, Cause_End, Effect_Start, Effect_End, Sentence*
- trial and training sample can be concatenated and contain the following information in headers::
  - Index: the ID of the text section
  - Text: Passage extracted from the 2019 financial news
  - Cause: Sub-string referencing the cause of an event (event or related object included)
  - Effect: Sub-string referencing the effect of the cause
  - Offset_Sentence2 : Offset of the start of the second sentence in the block
  - Offset_Sentence3: Offset of the start of the third sentence in the block
  - Cause_Start: (Offset) start of the sub-string annotated as cause
  - Cause_End: (Offset) end of the sub-string annotated as cause
  - Effect_Start: (Offset) start of the sub-string annotated as effect
  - Effect_End: (Offset) end of the sub-string annotated as effect
  - Sentence: Text with cause(e1) and effect (e2) offset tags.
- The blind test sample will display the same distribution as trial and training sets.

5.1 Task 2 - Evaluation Metrics

A baseline and an evaluation script for this task will be released with the blind test data. The evaluation metrics will be based on the number of exact matches and a weighted average F1 score, where the F1 score of each class is balanced by the number of individuals in each class.

5.2 Task 2 - Expected Submission

The participants are expected to submit their results by augmenting the blind test dataset with two columns containing their predicted results, in a csv file. Their contribution will be scored from the number of exact matches and the average F1 metric provided in 5 Task 2 Evaluation Metrics.

- The blind test data will be provided in a csv file with headers: *Index; Text; Offset_Sentence2; Offset_Sentence3*

Index	Text	Cause	Effect
1	Boussard Gavaudan Investment Management LLP bought a new position in shares of GENFIT S A/ADR in the second quarter worth about $199,000. Morgan Stanley increased its stake in shares of GENFIT S A/ADR by 24.4% in the second quarter.Morgan Stanley now owns 10,700 shares of the company's stock worth $211,000 after purchasing an additional 2,100 shares during the period.	Morgan Stanley increased its stake in shares of GENFIT S A/ADR by 24.4% in the second quarter	Morgan Stanley now owns 10,700 shares of the company's stock worth $211,000 after purchasing an additional 2,100 shares during the period.
2	Zhao found himself 60 million yuan indebted after losing 9,000 BTC in a single day (February 10, 2014)	losing 9,000 BTC in a single day (February 10, 2014)	Zhao found himself 60 million yuan indebted

Table 2: Task 2 - Training Data Main Strings Examples

- The expected results is to be provided by the participants in a csv file with headers: *Index; Text;* Offset_Sentence2, Offset_Sentence3; **Cause; Effect**

605.1 ; Year-to-date revenue advanced 72% to almost CA$26.7 million. (Source: EnWave Corp, August 29, 2019, op cit.) The revenue growth was fueled by increased orders for REV equipment by cannabis companies, and ongoing growth in Moon Cheese. The company continued to be cash-flow-positive, with adjusted earnings before interest, tax, depreciation, and amortization (EBITDA) of CA$139,000. ; 60 ; 108 ; **The revenue growth was fueled by increased orders for REV equipment by cannabis companies, and ongoing growth in Moon Cheese. ; Year-to-date revenue advanced 72% to almost CA$26.7 million.**

605.2 ; Year-to-date revenue advanced 72% to almost CA$26.7 million. (Source: EnWave Corp, August 29, 2019, op cit.) The revenue growth was fueled by increased orders for REV equipment by cannabis companies, and ongoing growth in Moon Cheese. The company continued to be cash-flow-positive, with adjusted earnings before interest, tax, depreciation, and amortization (EBITDA) of CA$139,000. ; 60 ; 108 ; **The revenue growth was fueled by increased orders for REV equipment by cannabis companies, and ongoing growth in Moon Cheese. ; The company continued to be cash-flow-positive, with adjusted earnings before interest, tax, depreciation, and amortization (EBITDA) of CA$139,000.**

752 ; The amendment is said to be government's move to uplift the slowing economy of the country. Few major changes brought in by the amendment are: A Reduced tax rate for the domestic companies not availing any tax exemptions A new provision, Section 115BAA has been added in the Income Tax Act, 1961 (IT Act) which specifies that income tax payable in respect of the total income of the domestic company shall be computed at the rate of 22% if the total income of the company has been computed: i. without any deductions allowed under provisions of IT Act ii. ; 91 ; NaN ; **The amendment is said to be government's move to uplift the slowing economy of the country. ; Few major changes brought in by the amendment are: A Reduced tax rate for the domestic companies not availing any tax exemptions A new provision, Section 115BAA has been added in the Income Tax Act, 1961 (IT Act) which specifies that income tax payable in respect of the total income of the domestic company shall be computed at the rate of 22% if the total income of the company has been computed: i. without any deductions allowed under provisions of IT Act ii.**

6 Contributions

The participants to this task will access the data after registering, and thereby pledge to contribute to the workshop by submitting an experiment paper. Participant can register to this shared task by filling this form to get access to the datasets. For any question please contact the organisers at fin.causal.task@gmail.com

References

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

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.