The Financial Document Causality Detection Shared Task (FinCausal 2023)

Antonio Moreno-Sandoval
Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
Madrid, Spain
antonio.msandoval@uam.es

Jordi Porta-Zamorano
Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
Madrid, Spain
jordi.porta@uam.es

Blanca Carbajo-Coronado
Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
Madrid, Spain
blanca.carbajo@uam.es

Doaa Samy
Department of Spanish
Cairo University
Cairo, Egypt
doaasamy@cu.edu.eg

Dominique Mariko
Machine Learning Lab
Yseop
Paris, France

Mahmoud El-Haj
UCREL
Lancaster University
Lancaster, UK
m.el-haj@lancaster.ac.uk

Abstract—We introduce the FinCausal 2023 Shared Task on Causality Detection in Financial Documents and the corresponding FinCausal dataset. This paper also provides insights into the participating systems and their outcomes. The primary objective of this task is to identify whether an object, event or sequence of events can be considered the cause of a preceding event (the effect). This year, we presented two subtasks, one in English and another in Spanish. In both subtasks, participants were tasked with pinpointing, within causal sentences, the elements that pertained to the cause and those that related to the effect. We received system runs from five teams for the English subtask and three teams for the Spanish subtask. FinCausal 2023 is affiliated with the 5th Financial Narrative Processing Workshop (FNP 2023), hosted at IEEE BigData 2023 in Sorrento, Italy.

Index Terms—causality detection, financial documents, NLP

I. INTRODUCTION

Grasping the concept of causality plays a pivotal role in examining decision-making processes. Financial analysis relies on factual data and understanding the underlying dynamics that give rise to these facts. But, while data provides the foundation of information, comprehending the processes leading to these facts proves indispensable. With this objective in mind, we have organised the Financial Document Causality Detection Task.

The task's purpose is to foster the capacity to elucidate, from external sources, the reasons behind transformations within the financial landscape. This effort serves as a precursor to generating precise and meaningful summaries of financial narratives. The task's primary objective is to assess which events or sequences of events can trigger alterations in a financial entity or unfold a significant occurrence, all within a specified context.

Participants were tasked with identifying, within causal sentences, the elements that pertained to the cause and those related to the effect in two different subtasks, one in English and another in Spanish. Notably, this marks the first year of introducing a Spanish subtask.

II. THE SUBTASKS

A causal relationship denotes the establishment of a causal link between a cause and its ensuing effect, signifying the interdependence of two events in which one serves as the trigger for the other. In essence, there exist two fundamental categories of causality:

1) The justification of a statement. For instance: "This is my final report, as I have assumed the position of Chairman of the Commission since January 24, 2019."
2) The rationale explaining an outcome. For example: "In Spain, turnover grew by 10.8% to 224.9 million euros, due to an increase in cement volume accompanied by a more moderate increase in selling prices."

At FinCausal, we are interested in both these causal types, which may be agents or facts. The effects, in this case, can either manifest as quantified facts or remain non-quantified.

For the English and Spanish tasks, participants can use any method they see fit (pattern matching, corpus linguistics, entity relationship models, deep learning methods) to identify the causes and effects.

A. The English subtask

This shared task focuses on determining causality associated with a quantified fact. An event is the emergence of a new object or context regarding a previous situation. So, the task will emphasise the detection of causality associated with the transformation of financial objects embedded in quantified facts. Participants were provided with a sample of text blocks extracted from financial news, annotated by two linguists.

Cause and Effect Detection. This task is a relation detection task. The aim is to identify the causal elements and the consequential ones in a causal sentence or text block (see Table I). Only one causal element and one effect are expected in each example. However, some examples can have more than one interpretation; see section III-A.

B. The Spanish subtask

This shared task focuses on determining causality associated with both events or quantified facts. For this task, a cause can be the justification of a statement or the reason that explains a result. Participants will receive a sample of paragraphs extracted from Spanish financial annual reports, labelled through inter-annotator agreement.

Cause and Effect Detection. This task is also a relation detection task. The aim is to identify, in a paragraph, the causal elements and the consequential ones (see Table II). Only one causal element and one effect are expected in each paragraph. In the Spanish dataset, there are no examples with several interpretations, unlike the English dataset. This is a decision of the task organisers, explained in section III-B.

III. DATASETS

The data has been procured from diverse sources. The English subtask encompasses financial news from various websites, while the excerpts are sourced from annual reports in the Spanish subtask. Consequently, despite both sources falling within the overarching domain of finance, they represent two distinct discursive subgenres, namely news and reports. The primary aim of this task is to examine whether the expression of causality exhibits variations based on subgenre and language within the same domain.

Unique dataset format for both languages. The files are delivered in UTF-8 plain text, in CSV format, where there are 4 columns per line, separated by ';':

```
Index;Text;Cause;Effect
```

Index: the ID of the example extracted from the document.
Text: the text segment containing one or more sentences with at least one causal relationship.
Cause: the text fragment containing the event that triggers the effect.
Effect: the text fragment containing the result triggered by a causal event.

The order of events can be interchangeable: cause ; effect, or effect ; cause, as in examples (1) and (2):

(1)   ⟨cause⟩Thanks to this measure⟨/cause⟩, ⟨effect⟩we have improved the density of such shipments by 1.5% over the previous year⟨/effect⟩.

(2)   ⟨effect⟩Monitoring of preventive measures to be implemented⟨/effect⟩ ⟨cause⟩as a result of the risk assessment and quarterly controls⟨/cause⟩.

The cause segment and the effect segment can be found in the same sentence —as in (1) and (2)— or in different sentences, as in example (3):

(3)   ⟨cause⟩Bankia is committed to sustainability in its business model and works towards growth with full respect for the environment.⟨/cause⟩ ⟨effect⟩Therefore, it integrates environmental management into the organization's decision-making process.⟨/effect⟩

A limitation of the annotation format is that it does not allow representing discontinuous segments or chained segments such as:

(4)   ⟨cause⟩ ... ⟨/cause⟩ ... ⟨effect⟩ ... ⟨/effect⟩ ... ⟨cause⟩ ... ⟨/cause⟩

A. English FinCausal dataset

The data is derived from a corpus of financial news articles collected by Qwam. The original raw corpus comprises a collection of HTML pages, each corresponding to daily information retrieved from financial news feeds. While these news articles provide insights into the financial landscape, they may also encompass information pertinent to politics, microeconomics, or other subjects deemed pertinent in financial information.

In the three previous editions of FinCausal (2020, 2021, and 2022), new examples have been added to the English dataset, which are summarised in Table III. Details of the compilation can be found in [1], [2] and [3].

As in previous editions of FinCausal, complex texts can have several (2 or more) interpretations. Typically, one cause can have two effects, or two causes can produce one effect. Those texts are identified by the IDs XX.XX.0, XX.XX.1, XX.XX.2... The training dataset contains 253 complex examples of this type (8.57% of the total). The test set has 24 complex examples (5%).

B. Spanish FinCausal dataset

The dataset has been sourced from a corpus of Spanish financial annual reports from 2014 to 2018, the FinT-esp corpus [4]. Participants were provided with a sample of text blocks extracted from the corpus, consisting of a paragraph with at least a causal event. Table IV details the composition and distribution.

This newly created dataset has some similarities and differences with respect to its English counterpart.

On the one hand, it maintains the same format and system of marking cause and effect. On the other hand, it is not limited to causal segments containing quantified effects. Therefore, the scope of discursive analysis has been extended to consequences triggered by a cause. We have avoided complex segments with multiple interpretations to compensate for the added difficulty. The annotators have also avoided labelling relations of finality, concession or condition between two events, which sometimes appear in the English version of FinCausal.

However, the most critical difference is in the evaluation set, which comprises a background set and a proper test set. The latter is the one that has been manually annotated to evaluate the systems, while the former is composed of causal samples

TABLE I
CAUSE AND EFFECT DETECTION SAMPLES FOR ENGLISH

Text	Cause	Effect
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.
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 II
CAUSE AND EFFECT DETECTION SAMPLES FOR SPANISH

Text	Cause	Effect
El deterioro de activos financieros se incrementó en un 267,4%, debido a ciertos impactos negativos de la cartera de clientes mayoristas y a la actualización del escenario macroeconómico.	debido a ciertos impactos negativos de la cartera de clientes mayoristas y a la actualización del escenario macroeconómico.	El deterioro de activos financieros se incrementó en un 267,4%
El resultado atribuido creció un 10,8% hasta los 1.522 millones de euros, gracias al buen comportamiento de las comisiones y sobre todo a la significativa reducción de los gastos y a los menores saneamientos y provisiones.	gracias al buen comportamiento de las comisiones y sobre todo a la significativa reducción de los gastos y a los menores saneamientos y provisiones.	El resultado atribuido creció un 10,8% hasta los 1.522 millones de euros

TABLE III
ENGLISH DATASETS BY YEAR

Edition	Dataset	Examples
2020	training + test	2394
2021	training + test	2493
2022	training + test	3098
2023	training	2952
2023	test	480

TABLE IV
2023 SPANISH DATASET

Dataset	Examples
training	2000
test	540
background	2990
test + background (evaluation set)	3530

of a certain complexity. This background set aims to see how the systems approach their labelling, which will serve as a pre-annotation for manual review by linguists. It consists of almost 3,000 paragraphs that will be part of the new dataset for future shared tasks.

IV. COMPETITION, PARTICIPANTS AND SYSTEMS

Competition was hosted in Codalab[1] [5]. A total of 37 teams from 13 countries have preregistered. By continents, we have:

- Asia: India, China, Pakistan, and UEA (4 countries)
- Europe: France, Spain, UK, Germany, Italy, and Romania (6 countries)
- America: USA and Canada (2 countries)
- Oceania: Australia (1 country)

By organisation type, the majority (70%) are universities or research institutes, although 30% are companies. Finally, 30 teams registered in CodaLab and downloaded the datasets.

A. The English subtask participants

In the final phase, five teams submitted results from their systems:

- MB-IIITH (India): RoBERTa model with a linear layer on top and a Viterbi decoder.
- NLP-UNED (Spain): Description not provided.
- SSC-AI-RG (India): Retrieval Augmented Generation and k-few-shot prompting with OpenAI GPT-4.
- LTRC-IIITH (India): ChatGPT with chain-of-thought prompting.
- Uni-Bucharest (Romania): Description not provided.

B. The Spanish subtask participants

In the final phase, three teams submitted results from their systems:

- BBVA AI Factory (Spain): Fine-tuned Spanish RoBERTa-large BNE[2] model with a bidirectional LSTM layer added on top and trained with augmented data.
- NLP-UNED (Spain), Spain: Description not provided.
- SSC-AI-RG (India): Retrieval Augmented Generation and k-few-shot prompting with OpenAI GPT-4.

V. EVALUATION METRICS

To evaluate systems, predictions and references are first converted into label sequences of the form (-, -, C, C, -, E, E, -, -), where each position corresponds to a token and the C,

---

[1]https://codalab.lisn.upsaclay.fr/competitions/14596

[2]https://huggingface.co/BSC-LT/roberta-large-bne

TABLE V
FINCAUSAL 2023 ENGLISH RESULTS

Rank.	Team	F1	Prec.	Rec.	Exact
1	MB-IIIT Hyderabad	0.71	0.71	0.72	0.25
2	NLP-UNED	0.62	0.60	0.66	0.17
3	SSC-AI-RG	0.58	0.56	0.62	0.00
4	LTRC-IIIT Hyderabad	0.54	0.52	0.58	0.08
5	Uni-Bucharest	0.46	0.45	0.47	0.06

TABLE VI
FINCAUSAL 2023 SPANISH RESULTS

Rank.	Team	F1	Prec.	Rec.	Exact
1	BBVA AI Factory	0.91	0.92	0.92	0.56
2	NLP-UNED	0.89	0.89	0.90	0.69
3	SSC-AI-RG	0.67	0.67	0.68	0.26

E, and - labels indicate whether the token is in the cause, the effect, or neither.

Several metrics are then used to evaluate the performance of the systems on these label sequences. The exact match metric represents the percentage of examples for which the predicted label sequence exactly matches the reference label sequence. Other token-level metrics, such as weighted F1 score, precision, and recall, are calculated by considering the tokens in the predicted and reference cause and effect sequences. The weighted F1 score is a weighted average of the F1 scores of each class, where the number of samples in each class determines the weights. This means that weighted F1 gives more importance to the performance of the systems in classes with more examples. Weighted F1 is the metric used for ranking the systems.

Like in previous editions of FinCausal, complex texts within the English task may carry several (2 or more) interpretations, often with one cause leading to two or more effects or one effect resulting from two distinct causes. Such texts with numerous interpretations can be located as instances that share the same text segment, with identifiers like XX.XX.0, XX.XX.1, XX.XX.2, and so on. For each possible interpretation, the scoring program iteratively tries to find the best match in predicted examples using F1, removing matches from the list of remaining predictions for the next iteration.

VI. RESULTS AND DISCUSSION

The models used in this edition of FinCausal can be grouped into two families: transformer-based models (based on RoBERTa) and generative prompting-based models (based on GPT). It is worth noting that generative models with prompt engineering, despite being more sophisticated, obtain inferior results to other simpler systems.

When analysing Tables V and VI, it becomes evident that notable disparities exist in the outcomes of the tasks. These tasks in Spanish and English exhibit dissimilarities in terms of the textual genre, encompassing news and financial reports and the underlying phenomena they address. Furthermore, the noteworthy observation is that two teams, namely NLP-UNED and SSC-AI-RG, participated in both tasks and achieved superior results in the Spanish task compared to the English one while purportedly employing the same system. This fact provides empirical substantiation for the assertion that the tasks are distinct.

Tables VII and VIII show examples of errors in both tasks. We have selected cases where none of the systems have correctly annotated cause and effect. It is observed that the systems have problems selecting the appropriate sequence for complex and extensive fragments. It is found that there is a reversal of order between the cause and the effect or modification (by paraphrase) of the original segment. The most common error is that the segment proposed by the system is shorter or longer than the Gold Standard. The English sample shows a difficult case, where three systems agree on the errors and the remaining two systems "invent" a paraphrase, produced by those systems trained with GPT.

Subsequently, we will examine the factors contributing to this variation, encompassing both the dataset design and the resultant outcomes.

Concerning the task focus —as we said above— the English subtask is centred on detecting causes and effects, specifically when the effects are quantified. In contrast, the Spanish task is designed to detect all causes and effects, not necessarily limited to quantified effects.

Shifting the task's focus from quantified effects to any form of effect or consequence resulting from a cause has driven us to narrow down the spectrum of relationships between two events. Within the English dataset, most dependency relationships between two events are permissible, including conditional or hypothetical ones (such as 'if this happens, then that will happen'). In contrast, in the Spanish dataset, we have confined the annotated examples to relationships where one event unambiguously yields another event. Consequently, relationships characterised by purpose ('we did this to achieve that'), concatenated events ('this happened, and then that happened'), or simultaneous events ('while this was happening, that was happening'), where causality is not apparent, have been omitted from the dataset.

The transition towards a more delimited conception of causality between events, driven by alterations in focus and data sources, could suggest that the task is better defined, resulting in improved learning by the systems. Notably, the English dataset featured 8% of complex sentences with multiple interpretations, a characteristic absent in the Spanish dataset. This implies that the systems had to be equipped to analyse a single example from various perspectives.

Finally, the incorporation of a background dataset in Spanish, comprising numerous instances of intricate relationships, is part of our strategy to encompass these phenomena in the forthcoming edition of the shared task.

TABLE VII
EXAMPLES OF ERRORS IN THE SPANISH TASK (SYSTEMS ARE ANONYMISED)

System	Text	Cause	Effect
Reference 22	Dichas manifestaciones de futuro o previsiones no constituyen garantías de un futuro cumplimiento, encontrándose condicionadas por riesgos e incertidumbres y los resultados reales pueden diferir materialmente de los anticipados en las manifestaciones de futuro o previsiones como resultado de diversos factores.	como resultado de diversos factores	los resultados reales pueden diferir materialmente de los anticipados en las manifestaciones de futuro o previsiones.
Systems A and B		como resultado de diversos factores	Dichas manifestaciones de futuro o previsiones no constituyen garantías de un futuro cumplimiento, encontrándose condicionadas por riesgos e incertidumbres y los resultados reales pueden diferir materialmente de los anticipados en las manifestaciones de futuro o previsiones
System C		encontrándose condicionadas por riesgos e incertidumbres y los resultados reales pueden diferir materialmente de los anticipados en las manifestaciones de futuro o previsiones	Dichas manifestaciones de futuro o previsiones no constituyen garantías de un futuro cumplimiento
Reference 3481	A lo largo de todo el año 2017, desde el equipo Global hemos acompañado a los equipos locales de Brasil, Chile y Perú en la transformación de su Modelo de Atención, con el objetivo de mejorar la satisfacción de los clientes de este segmento. Bajo las premisas de la personalización, convergencia y una visión E2E, gracias a lo cual se han conseguido notables mejoras en la experiencia del servicio posventa.	A lo largo de todo el año 2017, desde el equipo Global hemos acompañado a los equipos locales de Brasil, Chile y Perú en la transformación de su Modelo de Atención, con el objetivo de mejorar la satisfacción de los clientes de este segmento. Bajo las premisas de la personalización, convergencia y una visión E2E	gracias a lo cual se han conseguido notables mejoras en la experiencia del servicio posventa
System A		las premisas de la personalización, convergencia y una visión E2E, gracias a	lo cual se han conseguido notables mejoras en la experiencia del servicio posventa
System B		gracias a lo cual se han conseguido notables mejoras en la experiencia del servicio posventa	Bajo las premisas de la personalización, convergencia y una visión E2E
System C		acompañado a los equipos locales de Brasil, Chile y Perú en la transformación de su Modelo de Atención, con el objetivo de mejorar la satisfacción de los clientes de este segmento	se han conseguido notables mejoras en la experiencia del servicio posventa

REFERENCES

[1] D. Mariko, H. Abi-Akl, E. Labidurie, S. Durfort, H. De Mazancourt, and M. El-Haj, "The Financial Document Causality Detection Shared Task (FinCausal 2020)," in *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*. Barcelona, Spain (Online): COLING, Dec. 2020, pp. 23–32. [Online]. Available: https://aclanthology.org/2020.fnp-1.3

[2] D. Mariko, H. A. Akl, E. Labidurie, S. Durfort, H. de Mazancourt, and M. El-Haj, "The Financial Document Causality Detection Shared Task (FinCausal 2021)," in *Proceedings of the 3rd Financial Narrative Processing Workshop*. Lancaster, United Kingdom: Association for Computational Linguistics, Sep. 2021, pp. 58–60. [Online]. Available: https://aclanthology.org/2021.fnp-1.10

[3] D. Mariko, H. Abi-Akl, K. Trottier, and M. El-Haj, "The Financial Causality Extraction Shared Task (FinCausal 2022)," in *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 105–107. [Online]. Available: https://aclanthology.org/2022.fnp-1.16

[4] A. Moreno Sandoval, A. Gisbert, and H. Montoro, "FinT-esp: a corpus of financial reports in Spanish," in *Multiperspectives in Analysis and Corpus Design*, M. Fuster-Márquez, C. Gregori-Signes, and J. S. Ruiz, Eds. Granada, Spain: Editorial Comares, 2020, pp. 89–102.

[5] A. Pavao, I. Guyon, A.-C. Letournel, D.-T. Tran, X. Baro, H. J. Escalante, S. Escalera, T. Thomas, and Z. Xu, "CodaLab Competitions: An Open Source Platform to Organize Scientific Challenges," *Journal of Machine Learning Research*, vol. 24, no. 198, pp. 1–6, 2023. [Online]. Available: http://jmlr.org/papers/v24/21-1436.html

TABLE VIII
EXAMPLES OF ERRORS IN THE ENGLISH TASK (SYSTEMS ARE ANONYMISED)

System	Text	Cause	Effect
Reference 123	Mr Huang, listed as one of the 400 wealthiest people in China, where he has extensive holdings, is fighting the freezing orders from the ATO, which claims he earned $172 million on undeclared income during the 2013, 2014 and 2015 financial years.	claims he earned $172 million on undeclared income during the 2013, 2014 and 2015 financial years.	Mr Huang, listed as one of the 400 wealthiest people in China, where he has extensive holdings, is fighting the freezing orders from the ATO,
System A		Mr Huang, listed as one of the 400 wealthiest people in China, where he has extensive holdings, is fighting the freezing orders from the ATO	claims he earned $ 172 million on undeclared income during the 2013, 2014 and 2015 financial years.
System B		he has extensive holdings, is fighting the freezing orders from the ATO	which claims he earned $172 million on undeclared income during the 2013, 2014 and 2015 financial years.
System C		Mr Huang, listed as one of the 400 wealthiest people in China, where he has extensive holdings, is fighting the freezing orders from the ATO	claims he earned $172 million on undeclared income during the 2013, 2014 and 2015 financial years.
System D		as manufacturers including Hero Moto-Corp and Honda Motorcycle continued to cut production to control inventory, which is at over 60 days.	Two-wheeler volumes fell 22.24% y-o-y to 15,14,196 units, also the sharpest-ever decline
System E		ATO claims he earned $172 million on undeclared income during the 2013, 2014 and 2015 financial years	Mr Huang is fighting the freezing orders from the ATO