Research the Textual Causal Relationship Extraction Methods for the Financial Sector

WANG Runchen[1] ，ZHANG Jian[1], SONG Wenguang[1,2], XUE Yuanhong[1]

1 School of Computer Science, Yangtze University, Jingzhou, Hubei, 434023, China;
2 School of Computer Science and Engineering, Guangdong Ocean University, Yangjiang, 529500, China.

DOI: https://doi.org/10.56293/IJMSSSR.2026.6013

IJMSSSR 2026
VOLUME 8
ISSUE 1 JANUARY - FEBRUARY **ISSN: 2582 – 0265**

Abstract: The inherent complexity of financial texts stems from their specialised nature, intricate structure, pervasive implicit causality, time-sensitivity, and data sparsity. Causal relationship extraction constitutes the core task in financial text analysis, proving crucial for risk warning, market forecasting, and investment decision-making. Firstly, from a linguistic perspective, we deconstruct explicit and implicit causal relationships, dissecting the constraints inherent in financial text characteristics. Secondly, we trace the technological evolution from pattern matching and machine learning to deep learning. We critically evaluate the trade-offs between pipeline and joint structures in deep learning, summarise the features of financial datasets such as FCause and FinRE, and conduct a systematic study of causal extraction methods for financial texts, exploring future development directions.

Keywords: Causal relationships; Natural language processing; Relation extraction

1. Introduction

Causal relations typically denote grammatical components that logically express causal connections and serve as connectors; moreover, they are understood as a category representing the broader concept of causality. This is defined as a relationship between two states of affairs wherein the existence of one state leads to the occurrence of another. When describing relationships between entities and connections between events, causal relationships can facilitate deeper exploration of an issue's essence and provide clearer exposition of the relationships between variables or events. Extracting and analyzing causal relationships within text enables further investigation of causal linkages, thereby achieving the objective of predicting the outcomes of events[Error! Reference source not found.] .

With the advent of the big data era, the financial sector—as a data-intensive domain possessing vast repositories of information—urgently requires professionals to conduct in-depth mining and analysis. Research focused on financial texts has emerged as a prominent direction within the field of natural language processing (NLP)[Error! Reference source not found.].Particularly within financial news text data, the richness of causal relationships provides invaluable insights for understanding market dynamics and forecasting future trends. Extracting meaningful information, uncovering connections between events, and presenting these in a clear and unambiguous manner to financial decision-makers to effectively support investment choices remains a pressing challenge.

2. Principles of Causation Extraction

Causal relationship extraction constitutes one type of relation extraction task, frequently serving as a core component within information extraction. Causal relationships are categorised into explicit and implicit causal relationships, typically classified based on the presence of explicit causal connectives and whether they can be directly traced. Explicit causal relationships are directly expressed through causal connectives or syntactic structures, and can be further subdivided into four categories: single cause-single effect, single cause-multiple effects, multiple causes-single effect, and multiple causes-multiple effects. Implicit causal relationships, lacking explicit markers or involving complex semantic reasoning, require inference through contextual clues and external knowledge. An example is chain reactions in aviation safety reports, where equipment failures, human error, and weather conditions intertwine to cause accidents.

2.1 Classification of Causal Relationships

2.1.1 Explicit Causal Relationships

Explicit causality refers to a manner of directly and unambiguously revealing causal logic between events or states within linguistic expression. This is primarily achieved through the use of specific explicit linguistic markers or reliance on structured syntactic features, rendering the causal chain clearly visible without necessitating contextual cues or deep inferential reasoning by the language user. Its core characteristics lie in its directness and quantifiability. Directness manifests in the surface layer of language, typically containing fixed, recognisable causal signals; quantifiability implies that these markers and structures can be relatively easily identified, annotated, and statistically analysed within texts. This holds significant value for research in natural language processing, discourse analysis, and logical argumentation.

Specifically, explicit linguistic markers can be categorised into three primary types:

Causal conjunctions: These specifically link cause and effect clauses. For instance, words such as 'because', 'since', and 'as' introduce the cause, while 'therefore', 'consequently', 'hence','thus', 'henceforth', 'hence', and 'it follows' explicitly introduce the effect.

Causal adverbs/adverbial phrases: Typically positioned before the predicate or at the beginning of the consequent clause. Though not directly linking the clauses, they clearly and directly indicate the causal relationship. Examples include 'consequently', 'thus', 'therefore', 'hence', 'so', as well as 'resulting in', 'ultimately', and 'leading to'. These often modify the entire consequential event, highlighting its causal nature.

Causal punctuation: Certain punctuation marks convey causal relationships. The semicolon (;) frequently links two logically interconnected independent clauses, implying the latter may be the result or explanation of the former. The dash is also commonly used to introduce causal explanations or consequential clarifications, serving to supplement and emphasise causal connections. Through structural arrangement, these symbols indirectly yet effectively construct causal logical frameworks.

2.1.2 Implicit Causal Relationships

Implicit causal relationships, lacking explicit linguistic markers, require semantic inference, contextual association, and domain knowledge to identify. Characterised by semantic obscurity, cognitive complexity, and open-ended reasoning, they constitute a focal research area within natural language processing. Investigating implicit causality holds significant importance for enhancing the accuracy and depth of natural language understanding.

Within natural language processing, the identification and inference of implicit causal relationships present a challenging task. The semantic obscurity of such relationships renders them difficult to recognise directly, as they remain embedded within the text's meaning and context. Semantic reasoning and contextual association are required to deduce the implied causal connections. For instance, in the sentence 'He fell ill, so he did not attend school,' the causal relationship is explicit. Conversely, in the sentence 'He did not attend school because he felt unwell,' the causal relationship is more implicit, requiring contextual and domain-specific knowledge for inference. This identification process must transcend superficial semantics by constructing semantic network models to analyze the deep logical connections between concepts. Current mainstream approaches combine attention mechanisms to capture contextual dependencies, utilising knowledge graphs to inject domain-specific common sense.

2.2 Textual Feature Analysis in the Financial Sector

Financial texts typically contain a wealth of specialised terminology and industry abbreviations, while also addressing complex financial concepts such as 'leverage ratio', 'price-earnings ratio', 'short selling' and 'going long'. This presents additional challenges for natural language processing applied to financial texts. Concurrently, the linguistic expression in financial texts is frequently intricate, featuring lengthy sentences, nested syntax, and ambiguous formulations, which further complicates syntactic analysis and semantic comprehension for natural language processing models. These characteristics result in financial texts containing a substantial volume of implicit causal relationships, presenting considerable difficulties for causal relationship extraction within such materials.

Financial texts also exhibit pronounced time sensitivity, with causal relationships potentially shifting according to market conditions, policy adjustments, or economic cycles. Consequently, temporal factors must be thoroughly accounted for when constructing models. Conversely, data sparsity remains a significant challenge in financial texts, as annotated data for specific financial events is scarce, thereby limiting the effectiveness of supervised learning approaches. Financial texts originate from diverse sources, including news reports, research studies, company announcements, and social media. These varied sources exhibit significant differences in style, tone, structure, and information density, making unified modelling challenging. Consequently, causal identification in financial texts is typically divided into discourse-level or sentence-level analysis. Building upon this foundation, integrating domain knowledge with advanced natural language processing techniques enhances the model's adaptability and accuracy.

Given the aforementioned distinctive characteristics and unique challenges associated with causal relationship extraction in financial domain texts, this paper aims to provide a systematic review of research progress, key technologies, and future directions in Financial Causal Relationship Extraction (Financial CRE). Its core objective is to map the research landscape within this field, analyze the strengths and limitations of existing methods, and explore effective solutions addressing the particularities of financial texts, thereby offering guidance for subsequent research and practical applications.

3. Analysis of Datasets Related to the Financial Sector

Common Chinese datasets for relation extraction in the financial domain include FCause (Financial Causality Extraction Dataset)Error! Reference source not found.ChFinAnn[1], CCKS - NEC - Financial Domain Causal Event Feature Extraction DatasetError! Reference source not found. and the 2019 Financial News Dataset.

Among these, FCause is currently the most directly relevant and publicly available Chinese financial causality dataset, sourced from Chinese financial news outlets such as Sina Finance and East Money. East Money Information, and other Chinese financial news sources. The FinRE dataset constitutes a substantial Chinese financial relation extraction corpus, defining 44 types of financial domain relations including causal relationships, sourced from financial news and research reports. The ChFinAnn dataset, primarily designed for event extraction, inherently encompasses rich causal and sequential inter-event relationships, sourced from announcements of listed companies on the Shanghai and Shenzhen stock exchanges. CCKS - NEC - The training set for the financial causal event element extraction dataset comprises 4,000 causally annotated entries, including sample IDs, sentences containing causal events, original event texts, causal event elements, and causal event actions. The 2019 financial news dataset comprises 20,000 text entries including news headlines, content, and publication dates. It requires preprocessing tasks such as data cleaning and causal annotation.

Table 1. Financial Text Dataset

Name	Sample size	Tag	Whether to disclose	Applicable	Inclusion relationship
FCause	3000	Y	Y	Sentence level	causal relationship
FinRE	18000	N	Y	Sentence level	Multiple bidirectional relationships
ChFinAnn Dataset	35000	Y	N	Chapter-level	Multiple bidirectional relationships
CCKS-NEC	9500	Y	Y	Sentence level	Causal relationship
2019 Financial News Dataset	20000	N	Y	Sentence level	Multiple two-way relationships

4. Implementation of Textual Causal Relationship Extraction Methods in the Financial Domain

Text analysis tasks for financial texts are primarily applied in risk control, market surveillance, investment decision-making, and financial regulation. Consequently, research centered on financial texts focuses on event-cause extraction and text sentiment analysis. Current mainstream approaches to cause extraction fall into three categories: pattern-matching methods, machine learning-based methods, and deep learning-based methods.

4.1 Pattern-Matching-Based Approach

Early relationship extraction methods commonly employed pattern-matching approaches, utilising semantic features, lexical symbol features, and self-constraints to extract causal relationships through pattern matching. For instance, Blanco et al[2]. proposed a method for identifying the syntactic structure of 'verb phrase-clause' units linked by specfic relational words through pattern matching. Ma Bin et al[Error! Reference source not found.]., taking events as fundamental semantic units, analysed the semantic dependency relationships between events and their patterns of semantic dependency during evolution, thereby proposing an event relationship recognition method based on semantic dependency cues.A semi-automated framework based on pattern matching has been constructed to identify highly generalisable and broadly applicable causal patterns while effectively addressing linguistic ambiguities. This approach offers novel insights for subsequent research into causal relationship extraction utilising pattern matching and semantic resources.

Within the domain of financial texts, the core concept lies in fully leveraging the causal trigger words and sentence templates unique to the financial sector. This necessitates substantial involvement from domain experts to observe textual characteristics. Drawing upon linguistic knowledge and financial expertise, rules are employed to constrain the structural information within the text, enabling it to match corresponding rules and thereby extract relationships between entities. Consequently, this approach offers high interpretability, with transparent and easily comprehensible decision-making processes. It demonstrates considerable accuracy in identifying explicitly stated causal relationships within financial texts. However, the rules selected by domain experts remain specific to particular fields, rendering causal relationship extraction models built using this method non-generalisable.

4.2 Machine Learning-Based Methods

With the rise of machine learning and the advent of classifiers such as Support Vector Machines (SVMs) and Conditional Random Fields (CRFs), relation extraction can be viewed as a classification problem. Qiu J et al[4]. proposed a novel research perspective, redefining causality as a specific temporal relationship, thereby transforming the causality extraction problem into a time series labelling problem. Zhao S et al[5]. proposed a constrained latent naïve Bayes model for text causal relationship extraction. By calculating the tree-core similarity of sentences containing conjunctions, they obtained categorical features for causal conjunctions, treating the presence or absence of causal relationships within sentences as a binary classification task. Machine learning methods have to some extent reduced reliance on manual rules and improved generalisation capabilities, yet their performance remains severely constrained by the quality of feature engineering.

4.3 Deep Learning-Based Approaches

With the continuous advancement of deep learning, deep models can automatically learn hierarchical feature representations from raw text. Deep learning-based methods have gradually evolved into the primary techniques for event causation extraction, categorised into two structural approaches: The pipeline architecture employs two distinct models: the input text first passes through an event extraction model to yield causal event pairs, followed by a relation extraction model to derive causal relation event pairs. The joint architecture, conversely, shares model parameters across extraction tasks, enabling a single model to produce causal relation event pairs from the input text [Error! Reference source not found.].

4.3.1 Causal Relationship Extraction Based on Pipeline Architecture

The pipeline architecture decomposes the event causation extraction task into two or more sequential subtasks, most commonly event extraction and causal relationship identification. It offers advantages of modularity, high flexibility, and well-defined data requirements. However, the issue of error propagation persists. Wang Huan[Error! Reference source not found.] proposes an entity-relation extraction method tailored for financial texts. This approach employs the pre-trained model FinBERT to extract features from input financial texts. It incorporates a temporal lattice network to fuse word-level and character-level features, utilising a character-level attention mechanism to merge these features. This methodology addresses the challenge of accurately extracting semantic features from financial texts. Nayak et al[6]. conceptualised the overall framework as a pipeline structure, constructing a Transformer-based encoder-decoder architecture. This approach maps input text into an event representation space containing structured information such as events, their participants, and attributes for encoding. During decoding, it dynamically generates a complete causal outcome. The model's generative framework supports complex causal patterns including one-to-many and many-to-one relationships, thereby overcoming the limitations of traditional classification models. Cui S et al [7]. proposed a refined two-stage pipeline architecture that constructs an argument correlation graph for relation classification, enabling the discovery of cross-event argument dependencies and advancing event-level relation extraction to a finer granularity.

4.3.2 Causal Relationship Extraction Based on Joint Structures

Joint structures aim to integrate multiple subtasks, such as event extraction and causal relationship identification, into a unified model. Through parameter sharing and task interaction, they achieve end-to-end joint learning. This approach offers advantages such as avoiding error propagation and enabling complementary information. However, it retains the drawback of limited flexibility. Han R et al[8]. proposed a joint model for event extraction and temporal relation extraction. By sharing contextual embeddings and neural representation learners, it achieves complementary information exchange between tasks. Utilising integer linear programming for end-to-end joint optimisation, it avoids error propagation inherent in pipeline models. Yang B et al[9]. proposed a joint extraction approach for events and entities within document contexts. By constructing a joint probability distribution and introducing unary and binary feature functions, they performed fine-grained modelling of trigger words, entities, and their relationships. Constraints were applied to ensure the validity of the output structure, offering novel insights for document-level information extraction and advancing the application of joint learning in the field of natural language processing. Liu Suwen et al[Error! Reference source not found.]. proposed a joint learning model based on multi-task thinking, where binary relation extraction and unary feature recognition make decisions collaboratively. By sharing underlying vector representations across both tasks, the model employs long short-term memory (LSTM) networks and gating mechanisms to learn interactive representations between the tasks, thereby enabling classification predictions.

4.3.3 Summary

The evolution of causal relationship extraction has progressed from complete reliance on human expertise to feature-based approaches, ultimately advancing towards end-to-end automated feature learning and representation. This trajectory has progressively enhanced models' automation capabilities and generalisation performance.

This paper summarises the research findings mentioned, with the results presented in Table 2.

Table 2: Causal Identification Model

Type	Model	F-value	Dataset
Pattern-matching-based approach	Bagging with C4.5 decision trees	0.899	SemCor 2.1
	APCluster Clustering Algorithm	0.555	Unpublished dataset
	Semi-automatic causal vocabulary-syntactic pattern extraction model	--	Unpublished dataset
Machine learning-based methods	Double-layered cascaded CRF	0.853	Unpublished dataset
	Constrained Implicit NaiveBayes	--	EventStoryLine
Causal Relationship Extraction Based on Pipeline Architecture	BiLSTM+CNN+ATT	0.868	Financial Text Entity Relationship Extraction Dataset
	Encoder-Decoder	--	FinCausal2020 、 FinCausal2021

	Double grid marking	--	CCKS2021 Financial Causality Dataset
Causal Relationship Extraction Based on Joint Structures	Joint Event and Temporal Relationship Extraction Model	--	Unpublished dataset
	Event and Entity Joint Extraction Model	--	ACE 2005
	Joint Decision-Making Collaborative Learning Model	0.453	BV-C

5. Future and Outlook

With the rapid advancement of large language models, deep learning-based causal event extraction methods will continue to evolve towards higher accuracy and stronger generalisation capabilities. Larger-scale, more diverse pre-trained models will deliver enhanced contextual understanding and reasoning abilities; research into model lightweighting, knowledge augmentation, and interpretability will become focal points, enabling causal extraction techniques to play a greater role in scenarios such as financial intelligence analysis, public sentiment monitoring, and investment decision-making.

Moreover, graph neural networks (GNNs) are increasingly being applied to enhance causal reasoning capabilities. By incorporating syntactic dependency or semantic graph structures, GNNs can explicitly model structural relationships between entities and events, demonstrating significant potential particularly in tasks involving cross-sentence, long-range, or implicit causal relationship extraction.

Funding: This work was supported by program for scientific research start-up funds of Guangdong Ocean University（YJR24010）. China University Research Innovation Fund: Research on campus Internship and Training Teaching Platform for Artificial Intelligence Talent Training(2023ZY010). Laboratory Safety Research Project of Guangdong Ocean University in 2023: Research on Intelligent System for Laboratory Safety Monitoring, Early Warning and Emergency Management; Yangjiang Campus Artificial Intelligence Interdisciplinary Practice Teaching Platform (PX-1302024001). Guangdong Province College Students' Innovation and Entrepreneurship Training Program (Project No. S202510566003).

Wang Runchen (February 2001-), female, master's student, research direction in artificial intelligence and machine learning, E-mail：1738535918@qq.com.
Corresponding author: Zhang Jian (1981.02-), male, associate professor, master's supervisor, research direction: data science and engineering, machine learning and artificial intelligence, E-mail： zhangjian0716@126.com.

References

1. Sharma, S., et al. (2023). FinRED: A Dataset for Relation Extraction in Financial Domain. arXiv preprint arXiv:2306.03736.
2. Blanco E, Castell N, Moldovan D I. Causal Relation Extraction[C]//Lrec. 2008, 66: 74.
3. Girju R, Moldovan D I. Text mining for causal relations[C]//FLAIRS. 2002, 2: 360-364.
4. Qiu J, Xu L, Zhai J, et al. Extracting causal relations from emergency cases based on conditional random fields[J]. Procedia computer science, 2017, 112: 1623-1632.
5. Zhao S, Liu T, Zhao S, et al. Event causality extraction based on connectives analysis[J]. Neurocomputing, 2016, 173: 1943-1950.
6. Nayak T, Sharma S, Butala Y, et al. A generative approach for financial causality extraction[C]//Companion Proceedings of the Web Conference 2022. 2022: 576-578.
7. Cui S, Sheng J, Cong X, et al. Event causality extraction with event argument correlations[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 2300-2312.
8. Han R, Ning Q, Peng N. Joint event and temporal relation extraction with shared representations and structured prediction[J]. arXiv preprint arXiv:1909.05360, 2019.
9. Yang B, Mitchell T. Joint extraction of events and entities within a document context[J]. arXiv preprint arXiv:1609.03632, 2016.