# 2SCE-4SL: A 2-Stage Causality Extraction Framework for Scientific Literature

Yujie Zhang[1], Rujiang Bai[1,*], Ling Kong[2] and Xiaoyue Wang[1]

[1]Institute of Information Management, Shandong University of Technology, Zibo, Shandong
[2]School of information management, Nanjing Agricultural University, Nanjing, Jiangsu

Abstract

Extracting causality from scientific literature is crucial for many knowledge-driven downstream tasks. This paper proposes a novel causality extraction framework for scientific literature, 2SCE-4SL(**2**-**S**tage **C**ausality **E**xtraction for **S**cientific **L**iterature). In this work, the process of causality extraction is divided into two stages: (1) In the first stage, terms and causal trigger words are parsed from causal sentences and form noisy causal triplets. (2) In the second stage, we design a Denoising AutoEncoder based on Transformer architecture to represent the causal sentences, which is used to learn the causal dependency and contextual information of sentences through causal trigger word tagging and noise elimination, as well as inject domain-specific knowledge. Finally, combining the causality structure of stage 1 and the causality representation model of stage 2, the true causal pairs are identified from the noisy causal triplets. We selected open access scientific literature dataset for experiments, and compared the effects of different disciplines, training data volume, document length, whether causality representation on results, and analyzed the reasons for such differences. The results of this study indicate that the average precision of 2SCE-4SL reaches 0.8146 and the average F1 is 0.8308, among which the full-text performance is the best and the average precision reaches 0.9420. We also verify the effectiveness of the causality representation in stage-2, two tasks demonstrate the architecture can capture the causal dependency of sentences, showing good performance. In summary, detailed contrast experiments and ablation experiments indicate that the 2SCE-4SL only needs a small amount of annotated data to have better performance and good domain adaptability.

**Keywords**

2SCE-4SL, causality extraction, causality representation, scientific literature mining

## 1. Introduction

Scientific literature is the main form of expressing innovative ideas. Nowadays, the growing number of academic papers provides rich materials for scientific research[1]. Mining useful elements from literature, such as entities, concepts, and terms, is of great significance for promoting scientific innovation. Therefore, Scientific Literature Mining(SLM) has become a field of concern for many interdisciplinary researchers[2, 3, 4].

Causality is the expression of the relationship between cause and effect. The Nobel Prize in Economics in 2021 was awarded to David Card, Joshua D. Angrist, and Guido W. Imbens for their outstanding contributions in causal inference[5]. In recent years, more and more articles about causality have emerged in computing and information science community, some research applies causality to recommendation system[6] and opinion mining[7], some used in the interpretability[8] and stability[9], some applied in causality extraction[10, 11] and NLP augmentation[12]. It indicates that the study of causality is an area worthy of further exploration.

Because causality can express higher-order logical relations between linguistics, discovering causality in scientific literature plays an important role in academic recommendation, knowledge discovery, intelligent reasoning, event abduction, causal inference, future scenario generation and so on. For example, two sentences shown in Fig. 1, *"<colorectal cancer, leads to, colonic obstruction>,<prognostic model, thus, predicting colorectal cancer>"* can be obtained by extracting causal pairs in the literature. These triplets can be constructed as Knowledge Graphs about causality (e.g., Event Logic Graph [13]), thereby improving the performance of automatic question answering systems. The triplets can also predict potential connections between nodes by link prediction on the network, thereby facilitating knowledge discovery and future scenario generation.

To the best of our knowledge, although the existing research on causality extraction has made good progress [14], most of them revolve around commonsense knowledge, and most of the corpus comes from general fields, such as news[15], web[16], social media[17], etc. At the same time, many existing methods rely on supervised extraction methods[10, 14], which is difficult to implement in large-scale unlabeled literature data. After reviewing the existing methods, we think the difficulties of causality extraction in scientific literature mainly include the

*3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2022), June 20-24, 2022, Cologne, Germany and Online*

*Corresponding author.
E M A I L : 1725zyj@gmail.com (Y. Zhang); brj@sdut.edu.cn (R. Bai);2019214002@njau.edu.cn (L. Kong); wangxy@sdut.edu.cn (X. Wang)

following three aspects:

- Most of the current study focuses on commonsense causality, there is few research on causality extraction for scientific literature;
- Due to long length, intricate linguistic features and lack of domain-specific knowledge in scientific literature, causality extraction is still difficult;
- Due to lack of large quantities of readily available annotation data, the performance of supervised causality extraction of scientific literature needs to be improved.

To solve these problems, we propose a novel framework, 2SCE-4SL(**2**-**S**tage **C**ausality **E**xtraction for **S**cientific **L**iterature), to extract causality from scientific publications. The framework consists of four parts:causality detection, causality collocation(stage 1), causality representation(stage 2), causality classification. In the first stage, causal trigger words and entities of sentences are identified and form to noisy causal triplets. In the second stage, a Denoising AutoEncoder based on Transformer architecture is designed for learning causal dependency and contextual information through causal trigger word tagging and noise elimination, while injecting domain-specific knowledge. Finally, actual causal pairs were identified from noise causal triplets in combination with causal expression structure of stage 1 and causal sentences representation model of stage 2.

The remainder of this paper is organized as follows: Section 2 introduces the related algorithms, datasets and applications of causality extraction. Section 3 introduces the framework structure and method design of 2SCE-4SL in detail. Section 4 introduces the dataset and empirical research, which mainly includes two parts of experiments: one is to verify the effectiveness of the causality representation architecture from 2 tasks; the other is to conduct comparative experiments and ablation experiments on 2SCE-4SL from 4 aspects; Finally, in the section 5, we summarize the findings, theoretical and practical implications, as well as discuss limitations and future directions.

## 2. Literature review

Causality Extraction (CE) is a sub-field of natural language processing. Although there have been many studies on this field, most of the algorithms and datasets of CE focus on commonsense, and their application scenarios include future scenario generation, event prediction, knowledge discovery, causal inference and so on. In this section, we reviewed exist CE methods, datasets and application scenarios, focusing on the research closely related to this work.

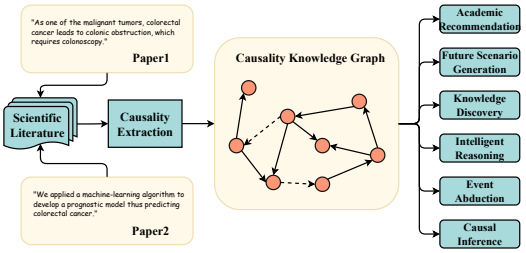


**Figure 1:** Application scenarios of causality extraction for scientific literature. Causal triplets can be extracted from scientific papers to construct knowledge graph about causality (e.g., Event Logic Graph), which can facilitate academic recommendation, future scenario generation, knowledge discovery, intelligent reasoning, event abduction, causal inference and other downstream tasks.

### 2.1. Causality extraction methods

Causality can be defined as the action of one event (cause) on another event (effect), where the latter is considered to be the effect of the former. Causality can be divided into explicit and implicit causality[14]. Causality extraction in this paper is limited to explicit causality. In natural language processing, CE can be implemented in a variety of ways, that is, through a variety of algorithms to automatically identify causality from text. According to different types of tasks,CE can be divided into : rule-based and machine learning-based.

CE methods based on manual rules (such as expert knowledge and pattern matching) identify causal relationships in texts by defining causal clues in advance or templates expressing causal structures[18]. This method built a large knowledge base and achieved good performance in CE. [19] proposed causal relationship extraction using cue phrase and word pairing probability. [20] proposed a method to mine the causal relationship of biomedical literature texts, they constructed two schemes: lexic-based causal term strength recognition, frequency-based causal strength and direction recognition. [21] used word vector mapping to extract causal relationships from literature. They used four types of verbs as candidates to form CE rules, and the results of causal extraction in Alzheimer's disease showed effectiveness. [22] developed COATIS for searching causal links in texts; [23] used pattern matching to identify causality explicitly expressed in a single sentence. In addition, dependency parsing and syntax tree parsing can effectively identify causal relationships in text. [15] proposed that the causal sentence pattern template < Pattern, Constraint, Priority> was constructed by rule constraint and syntactic parsing. This method extracted headlines of news articles in 150 years, the recall rate is 10% and the Precision is 70%.

Although the CE method of rule-based has high accuracy, the significant disadvantage is that it relies on

external rules, which requires a lot of manpower and time, and it is difficult to achieve multi-scenario generalization.

CE method based on machine learning makes up for the shortage of manual, with the deepening of neural network and deep learning greatly improve efficiency. CE methods based on machine learning can be divided into three types: text classification, relation extraction and sequence labeling. (1) Text classification method refers to classifying sentences according to whether the text contains causality or not[24, 25]. This method does not need to extract entities or events, but only needs to determine whether the text contains causality. It is suitable for those causal data that are difficult to extract events or entities from sentences. For example, [17] proposed a method based on text classification to detect causality from tweets. (2) Relationship extraction method is to judge whether the causal pair given in the text has a causal relationship, which is applicable to the data that is easy to identify the causal entity. This method is currently popular, and there are many available models, such as causality detection based on Bayesian algorithm [26], causality recognition driven by background knowledge[27], knowledy-driven CNN[28], etc. [29] proposed the extraction of time relationship and causality based on event network. They proposed an unsupervised event network representation structure, generated the causal relationship of triples, and then constructed a network to connect the relationship of events. (3) Sequence labeling method refers to mark the causal relationship label of each word in the text and then train the model to carry out generative extraction. These methods are basically end2end pipelines without too much manual intervention. [10] proposed a CE algorithm based on self-Attentive BilstM-CRF, which can achieve a high accuracy. [11] proposed a linguistically opposed approach to inform Bi-LSTM, which can also achieve good results.

On the one hand, although machine learning-based methods are efficient, they require a large amount of annotation data. On the other hand, although there are some standard datasets for CE, such as SemEval[30], Altlex[24], CEC[31], there are very few datasets about CE in scientific literature. It is urgent to establish perfect evaluation and datasets.

## 2.2. Application of causality extraction on scientific research

Application of CE in scientific research is mainly to promote knowledge discovery, academic recommendation, future scenario generation, event abduction and causal inference.

Extracting causal structure of triplets cloud facilitates discovery of new knowledge. [32] proposed CausalTriad, a framework for causal relationship discovery and hypothesis generation based on medical text data. In order to model the rules of causality transfer in medical texts, they divided the network composed of candidate pairs of causality into a large number of triadic structures, and then used the connection of text information and structural knowledge to mine medical causality. The experiment showd that CausalTraid is very efficient in discovering causality between sentences. Meanwhile, their team also proposed a fact-condition joint extraction pipeline[33] for scientific literature to identify scientific observations and research assumptions in scientific literature. In previous work, we conducted a preliminary study on causal discovery and knowledge linkage in the biomedicine[34]. We constructed causal knowledge network by extracting the causal triplets in the scientific literature, so as to analysis of scientific knowledge community and the prediction of potential medical knowledge.

In addition, the discovery of causal relationships in scientific research is beneficial to future scene generation as well as scientific event detection and prediction. [35] proposed a pretraining language model, EGE -RoBERTa, based on variational autoencoder to enhance the knowledge of the problem atlas, which utilizes an additional implicit variable to capture the necessary problem atlas knowledge. Experimental results showed that this model can improve the performance of abductive reasoning effectively compared with baseline method. [7] proposed to discover the causal background from political tweets and reveal the context of political opinions and news reports. They integrated OpenIE, open knowledge repository and deep neural network to extract meaningful tweet clauses and analyzed causal correlation,which show good results.

More importantly, with the development of causal inference in the field of artificial intelligence, the deep integration of causality and NLP can promote the interpretability and stability of scientific research. [12] and other 13 scholars reviewed the current direction of the combination of causality and NLP. One is that NLP integrates causal inference, and the other is that causal inference enhances NLP. In addition, the excavation of causal relationships in scientific literature plays an important role in promoting the understanding of academic activities. For example, [36] explored what makes science paper acceptable for publication, and they provided a method for detecting the confounding effects of scientific literature in order to generate causal explanations for the dynamic activities of academic research in a scientific collaboration model.

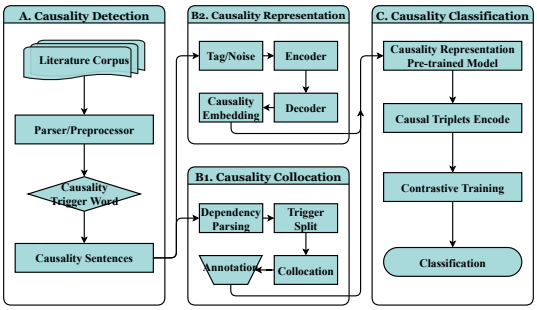


**Figure 2:** Framework of 2SCE-4SL. It consists of four parts: Causality Detection(A), Causality Collocation(B1, stage 1), Causality Representation(B2, stage 2), and Causality Classification(C).

# 3. Methodology

## 3.1. Overview

Given literature set $\mathbb{D}, \mathbb{D} = \{d_1, d_2, d_3, ..., d_i | (1 \leq i \leq N)\}$, the goal of this paper is to identify triplets $(c, v, e)$ that express explicit causality, with $c$ representing cause, $e$ representing effect, and $v$ representing causal verb. We will solve this problem in a classification framework.

The framework of 2SCE-SL is shown in Fig. 2, which consists of four parts: Causality Detection (Fig. 2. A), Causality Collocation (Fig. 2. B1), Causality Representation (Fig. 2. B2) and Causality Classification (Fig. 2. C). In causality detection, detecting set $\mathbb{S}$ expressing causality from $\mathbb{D}$, $\mathbb{S} = \{s_1, s_2, s_3, ..., s_n | \exists s_n \in \mathbb{D}\}$. In causality collocation(stage 1), parsing each causal sentence $s_n$, collocating the terms and the causal trigger word to noisy causal triplets $\mathbb{C}(\exists(c, v, e) \in \mathbb{C})$, then manually annotate a small amount of data. In causality representation (stage 2), $\mathbb{S}$ is represented by the AutoEncoder architecture to learn the causal dependency and semantic structure of sentences, then output the causal representation model $\boldsymbol{M}$. Finally, in the causality classification, combined with the causal expression structure $(c, v, e)$ of stage 1 and the causal sentence model $\boldsymbol{M}$ of stage 2, the true causal pairs will be identified from the noisy causal triplets, which can be defined as follow:

$$f((c, v, e)) = \begin{cases} y = 1 & , \text{ causality} \\ y = 0 & , \text{ not causality} \end{cases} \quad (1)$$

Given an arbitrary causal triplet $(c, v, e)$ ,if label is 1, indicating true causality; otherwise false causality. Each section will described in detail below.

## 3.2. Causality detection

As the premise of causality extraction, the purpose of causality detection is to identify sentences that express explicit causality in the literature. Taking sentences as a unit can not only retain the complete semantics to facilitate the feature representation of causal structures, but also provides rich conditions for further fine-grained parsing.

Here, we divide the causality detection of scientific literature into three steps:

- Firstly, preprocess the literature corpus;
- Then, define trigger words that express causality;
- Finally, the sentences containing causal trigger words are identified.

In the selection of causal trigger words, we referred to the previous work[17, 37, 38], defining 81 causal cues such as *"lead to, result in, because"*. Part of the causal trigger words are shown in table1.

## 3.3. Causality collocation

As the first stage of 2SCE-4SL, causality collocation aims to identify candidate causal triplets from causal sentences, which is one of the targets extracted in this paper. The method based on deep learning requires a lot of annotated data, so we construct a large number of causal pairs in the way of terms collocation, and then identify the true causal pairs in the causality classification part, which can be regarded as a part of few-shot learning.

We preliminarily define the structure of causality, which is represented by the triple form of $(c, v, e)$, where $c$ stands for cause,$e$ stands for effect, and $v$ stands for causal trigger word. This triplet structure of causality could facilitate direct application in many downstream tasks such as Fig. 1. Firstly, the causal trigger word $v_i$ is identified from the causal sentence. Then, $v_i$ is used as the boundary to identify the entities $c_i$ and $e_i$ on both sides. For any $\forall c_i$ and $e_i (c_i \neq e_i)$, ergodic combination is carried out and Cartesian product is calculated to construct the triplets set $\mathbb{C}$, which can be expressed as follows:

$$\begin{aligned} \mathbb{C} &= \{c_1, \ldots, c_i\} \times \{v_i\} \times \{e_1, \ldots, e_i\} \\ &= \{(c_1, v_i, e_1), (c_1, v_i, e_i), \ldots \mid c_i \neq \mathrm{e_i}, \mathrm{c_i} \& \mathrm{e_i} \in \mathbb{S}\} \end{aligned} \quad (2)$$

As shown in Fig. 3, input a sentence *"As one of the malignant tumors, colorectal cancer leads to colonic obstruction, which requires colonoscopy."*, using ScispaCy [39] to perform dependency parsing, and identify the concepts, terms or phrases. Taking the trigger word *"leads to"* as the boundary, causal triplets *"<the malignant tumors, leads to, colonic obstruction>, <colorectal cancer, leads to, colonic obstruction>, <the malignant tumors, leads to, colonoscopy>, <colorectal cancer, leads to, colonoscopy>"* can be parsed from sentence.

Obviously, these triplets contain lots of noise. *"<colorectal cancer, lead to, colonoscopy>"*, for example, does

**Table 1**
Part of the causal trigger words.

| | | | | | |
|---|---|---|---|---|---|
| arouse | caused by | give rise to | inasmuch as | owing to | stem from |
| beacuse | coz of | have effect on | induce | provoke | that's why |
| beacuse of | elicit | hence | lead | result from | therefore |
| bring about | engender | if, then | lead to | result in | thus |
| by reason that | evoked | in consequence of | on account of | so that | ... |

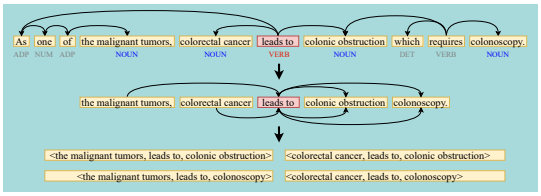


**Figure 3:** Example of causality collocation. Input a sentence,*"As one of the malignant tumors, colorectal cancer leads to colonic obstruction, which requires colonoscopy."*, using ScispaCy to perform dependency parsing, and identify the concepts, terms or phrases.Taking the trigger word *"leads to"* as the boundary, causal triplets *"<the malignant tumors, leads to, colonic obstruction>, <colorectal cancer, leads to, colonic obstruction>, <the malignant tumors, leads to, colonoscopy>, <colorectal cancer, leads to, colonoscopy>"* can be parsed from sentence.

not express causality in the paper. The goal of the following processes are to filter these noisy triplets. We manually annotated part of the data as training data to identify true causal pairs in classification section.

## 3.4. Causality representation

As the second stage of 2SCE-4SL, causality representation is a very important component of this paper, which aims at characteristic representation of causality sentences so as to learn causal dependency inside sentences and contextual knowledge of specific domain.

A prominent feature of scientific literature that distinguishes it from commonsense corpus is that it contains professional knowledge. For example, Computer Science paper often contains professional concepts and terms. In addition, the expressions of causal sentences also have specific linguistic features. The entities on the same side of the triggering word $v_i$ are more continuous sequence structures, while the entities on the other side of $v_i$ are not only more discrete in continuity, but also farther apart in vector coordinates. In order to accurately identify positive causal pairs from noise, it is necessary to learn the causal dependency and semantic information inside these sentences.

Here, we constructed an encoder-decoder architecture network based on Transformer model[40]. Its composition is shown in Fig. 4, which can be regarded as an AutoEncoder for learning causal structure. Inputting causal sentence, the position information of $v_i$ is tagged in Encoder based on Denoising AutoEncoder(DAE)[41], then in Decoder, randomly add some noise to the sentence, and encode the fixed length vector, as well as randomly mask the words around $v_i$. The goal of training is to recover the embeddeding representation of the original causal sentence from the noise data and learn its semantic structural information.

Specifically, tagging the position of trigger word during input aims to distinguish the cause and effect of sentences. Randomly added disturbances (delete, add, exchange, etc.) and masks aim to minimize the loss function, restore the lost information compressed around $v_i$, and improve the robustness of causality representation. Our training objective function can be formalized as follows:

$$\begin{aligned} J_{DAE}(\theta) &= \mathbb{E}_{s\sim\mathbb{S}}\left[\log P_\theta(s \mid \widetilde{S})\right] \\ &= \mathbb{E}_{s\sim\mathbb{S}}\left[\sum_{t=1}^{l}\log P_\theta\left(s_t \mid \widetilde{S}\right)\right] \\ &= \mathbb{E}_{s\sim\mathbb{S}}\left[\sum_{t=1}^{l}\log\frac{\exp\left(h_t^T e_t\right)}{\sum_{i=1}^{N}\exp\left(h_t^T e_i\right)}\right] \end{aligned} \tag{3}$$

Where $\mathbb{S}$ is the corpus set of all causality sentences,$s_l$ is the input sentence with length $l$, and $\widetilde{S}$ is the sentence with noise added to $s$. $e_t$ is the word embedding representation of $s_t$,$N$ is vocabulary size, $h_t$ is the hidden state of the encoder's $t$ step output. We use cross-attention as part of Decoder, which can be formalized as follows:

$$\begin{aligned} H^{(k)} &= \text{Attention}\left(H^{(k-1)}, \left[s^T\right], \left[s^T\right]\right) \\ \text{Attention}\left(Q,K,V\right) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \end{aligned} \tag{4}$$

$H^{(k)}$ is the hidden state within $t$ steps of $k$ layer in Decoder, and $d$ is the dimension of causal sentence vector; $\left[s^T\right] \in \mathbb{R}^{l\times d}$ is a sentence vector output by Encoder. No matter which layer of cross-attention,$K$ and $V$ will always be $\left[s^T\right]$. The purpose of this design is to manually add a bottleneck to the model, make Encoder more accurate and increase the adaptability of the domain. This unsupervised network structure does not require annotation data, but only needs fine-tuning to encode the structure of causality sentences, so as to learn causal dependency and prior knowledge.

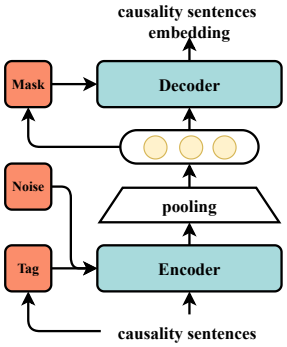


**Figure 4:** Architecture of causality representation.The architecture is an AutoEncoder based on Transformer model, which is composed of an Encoder and a Decoder.The Encoder tags the trigger word of a causal sentence when input, and then randomly adds disturbance information (deletion, addition, exchange, etc.), randomly masks some words before the Decoder. The training object is to recover the embedding representation of the original causal sentence from the noise data and learn semantic structural information.

## 3.5. Causality classification

Causality classification is the last composition of 2SCE-4SL, which combines the triplet structure of the stage 1 and the hidden state vector of the stage 2. The positive causal pairs will be identified from noisy causal triplets based on fine-tuning model $\boldsymbol{M}$.

Given any causal pair $\forall(c, v, e)$, the purpose is to determine whether causality, we convert this task into binary classification. In order to better solve the problems of noise interference and insufficient labeled data, contrastive learning method[42] is adopted. Dividing the training data into positive samples and negative samples, in which positive samples is any causal pair in the same category while negative samples is different pair. The constructed positive and negative samples are introduced into the stage 2 for encoding. For each encoded causal pair $s_t$, its objective is to identify the confidence degree of $s_t$ according to model parameters and features of causal sentences. The loss function during training can be formalized as follows:

$$\begin{aligned}\mathbb{C}ost\left(h_\theta\left(s_t\right), y\right) = &- y_i \log\left(h_\theta\left(s_t\right)\right) - \left(1 - y_i\right) \\ &\log\left(1 - \left(h_\theta\left(s_t\right)\right)\right)\end{aligned} \tag{5}$$

When $y = 1$, that's the positive sample, and if $h_\theta\left(s_t\right) = 1$, then $\mathbb{C}ost = 0$ for this sample alone, that means that this sample is completely accurate. If all samples are correctly predicted, the total $\mathbb{C}ost$ will approach 0. But if the probability $h_\theta\left(s_t\right) = 0$ is predicted at this time, then $\mathbb{C}ost \to \infty$. By iteratively learning the semantic relationship between causal pairs, the parameter $\theta$ of hidden $h_t$ was updated to reduce the noise of pseudo causal pairs and identify the true causal pairs.

# 4. Experiment

## 4.1. Dataset and evaluation

### 4.1.1. Dataset

We chose S2ORC[43] as our experimental data, which is a common English corpus for NLP and text mining research of scientific papers developed by Allen Institute for AI. The S2ORC is collected from open access platforms such as MAG,arXiv, PubMed and stored in a structured form. The dataset contains 81 million papers in 20 disciplines such as Computer Science, Material Science, Economics, Medicine, etc.

In this work, we took Computer Science(CS) and Medicine(Med) as examples to extract causality. These two fields are the most popular subjects at present, both of which belong to interdisciplinary subjects. Moreover, choosing different disciplines can verify the robustness of the 2SCE-4SL. Because the S2ORC is very large (over 12.7 million full-text papers), we randomly sampled some papers and selected the title, abstract and full-text as extraction objects respectively. CS and Med each collected 100,000 papers for a total of 200,000 papers.

### 4.1.2. Evaluation

We designed a architecture for representing causality sentences in section 3.4, but how to measure the quality and validity of causality representation? Two experiments will be evaluated in section 4.2.2:

- Whether the architecture captures causal dependency of sentences. Theoretically, if causal dependency can be learned, the similarity of causal pairs of same label will be higher, and that of the dissimilar label will be lower. Therefore, we will calculate the similarity between labeled causal pairs to measure causal dependency.
- Whether the architecture learns the semantic information of causal sentences. We will use the Semantic Textual Similarity (STS) task for evaluation by randomly drawing sentences and calculating their similarity.

More importantly, the necessary contrast experiments and ablation experiments are key to verify the effectiveness of 2SCE-4SL. To compare the effect, SciBert[44], a pre training model based on scientific papers, will be compared with 2SCE-4SL as baseline method for CE. Specifically, we will compare and analyze the four aspects in 4.2.3:

- Whether there are differences in causality extraction among different disciplines (CS, Med);
- Whether different amounts of training data affect the results;

**Table 2**
Causality sentences on CS and Med.

| | CS | Med |
|---|---|---|
| Title | 160 | 2526 |
| Abstract | 4130 | 9251 |
| Full_Text | 113063 | 134588 |
| Sum | 117353 | 146365 |

- The effect of whether causality representation (stage 2) on the results;
- Whether there are differences in the performance results of different document lengths (title, abstract, full-text).

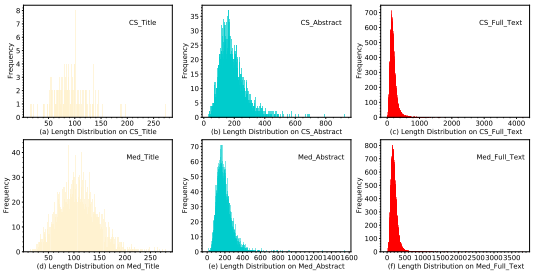


**Figure 5:** Length distribution of causality sentences. The length distribution of causal sentences varies with different types of documents, with the title being the shortest and the full text the longest, and gradually approaching the extreme long-tail distribution with the increase of sentence length.

## 4.2. Results

### 4.2.1. Causality detection and collocation on scientific literature

We parsed JSON-structured S2ORC, identify CS and Med literature (including multi-label discipline literature), and parse the title, abstract, and full-text. Then, causal sentences were detected based on pre-defined causal trigger words, including 117,353 in Computer Science and 146,365 in Medical Science, as shown in table 2.

There are differences in the distribution of different types of documents. As shown in Fig. 5, it can be found that compared with the distribution of abstract, the sentence length distribution of titles is shorter, with the overall distribution around 100. However, the causal sentences in the whole text are longer and follow the typical long-tail distribution. The distribution of sentence length may produce different features in the representation of causality, which will have different effects on the subsequent classification effect of the causal triplets.

In the stage 1, based on the detected causal sentences, syntactic analysis and causality collocation were carried out for these sentences. We parsed it in ScispaCy, loading the *en-core-sci-scibert* model to parse out entities, proper nouns, or combinations. After collocation, three annotators in the professional field will search for the corresponding statement from the article according to the index. If true causal pair, it will be marked as 1; otherwise, marked as 0. Only when at least two annotations are consistent, it will be valid. In the titles, abstracts and full -texts of the two disciplines, 400 are marked respectively, and 2400 causal pairs are labeled in total. These annotated data will be used for few-shot training.

### 4.2.2. Causality representation on scientific literature

In the stage 2, the pre-training model was used for fine-tuning. Unlabeled causality sentences were put into the model, SciBERT[44] was used as the initial model, and then the position information of the causal trigger words was tagged. These input causality sentences are automatically encoded by adding noise functions through the *DenoisingAutoEncoderLoss* function. The added noise includes delete, add, swap, and mask. As the representation training cost a lot of resources,we adopted the method of parallel training. 10000 pieces of data were input for each batch of different subjects, batch size set to 8, epochs set 10, weight decay 0 and learning-rate 3e-5.

(1) Evaluation for causal dependency

The similarity of the causal pairs of same label and different label were calculated respectively, the results are shown in Fig. 6.When label both 1, the similarity of causal triplets is about 0.3(Fig. 6 (a)) and 0.55(Fig. 6 (b)), and when labels are 0, the average similarity is 0.2. When the positive and negative samples are mixed, the average similarity is only less than 0.2. It can be found that the similarity between positive samples is the highest, while the similarity of mixed labels is not as good as that of the same labels on the whole, and the dispersion degree between their causal pairs is more significant, indicating that the model can capture the structure between true causal triplets. Learn about causal dependency from these causal sentences. It is worth noting that the similarity between the positive samples of CS is low, only about 0.3, and the upper and lower limits are further, which may be due to the annotation quality of these samples needs to be improved, and cannot be well distinguished from negative samples, which need be further optimized in the future.

(2) Evaluation for semantic information

To assess the ability of the causality representation architecture to capture semantic information, we measured its performance on the STS task. Inputting a causal sentence randomly *"However, small perturbations of the CiteSeer data lead to significant changes to most of the clusters."*, return the 5 sentences closest to it. The similarity of the first sentence *"Removing such large groups*

**Table 3**
Metrics on different subjects and training data

| Subject→ | CS | | | Med | | |
|---|---|---|---|---|---|---|
| Size↓ / Metric→ | P | R | F1 | P | R | F1 |
| 300 | 0.6923 | 0.8780 | 0.7742 | 0.8367 | 0.7885 | 0.8119 |
| 600 | 0.8696 | 0.8163 | 0.8421 | 0.7927 | 0.8500 | 0.8204 |
| 1200 | 0.8811 | 0.9056 | 0.8932 | 0.8152 | 0.8721 | 0.8427 |

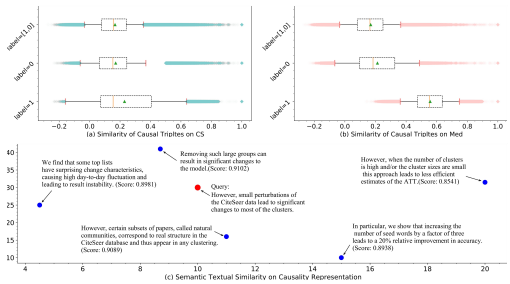


**Figure 6:** Evaluation for the causal dependency and semantic information.(a) and (b) are the sample similarity of CS and Med respectively. When both labels are 1, they have higher similarity, indicating that causal dependency of sentences can be captured. (c) is semantic search for causality representation. Input a paragraph to query the closest result, 5 returned records indicate that semantic information of causal sentences can be captured.

*can result in significant changes to the model"* is 0.9102, which shows that it has a good performance. The other returned causal sentences have certain semantic similarity more or less, indicating that the structure can capture the semantic features of sentences. Although there are still some differences in semantic search, we believe that the effect will be better when the corpus is increased.

#### 4.2.3. Causality classification on scientific literature

(1) Comparison of different data volumes

Firstly, the classification metrics of CS and Med causal pairs were compared, training data of 300,600 and 1200 were set respectively and split 7:3,train-loss set as *cosine similarity loss*, iterative set 10, results are shown in table 3.

Although belonging to few-shot training, it is similar to supervised learning, the increase of training data will improve the precision and F1 of the model accordingly. Therefore, in the future semi-supervised learning can be adopted to obtain more training data in subsequent studies, and then the model can be trained twice to improve the capability of 2SCE-4SL by increasing the amount of data.

(2) Ablation experiment: whether causality representation

Does the causality representation of the stage 2 contribute to the outcome? To test this idea, we performed ablation experiments to compare the classification performance before and after stage 2. Without representation means that manually annotated data are directly put into pre-training models (such as BERT and SciBert) for training, and then classified after fine-tuning. Here, we take the original SciBert as the baseline of ablation experiment, and take the model after the representation of causality sentences in SciBert as the comparison. Other parameters are consistent with those before, and the training data is set to 1200. After 10 iterations, The results on CS and Med are shown in table 4. It can be found that the precision before representation are 0.8481 and 0.7368 respectively; After stage 2, precision were 0.8811 and 0.8152 respectively, indicating that increasing causality representation at stage 2 significantly improved the results.

When the causal sentences is represented, the position information tagged with the trigger word can clearly distinguish the cause and effect, and the position information is added into the logical structure when encoding to make the structure of the causal sentence clearer. The test datasets of CS and Med are mapped to 2D coordinates, as shown in Fig. 7, the results are consistent with table 4. After stage 2, the true causal pair is more concentrated, the overall distribution shows good separability, and the false causal pair is more discrete, which confirms the help of causality representation to classification. After encoder and decoder add noise, the causal sentence can restore the information of the sentence itself, obtain the implied parameters of the sentence, and obtain the feature extraction ability even in the case of small sample size. Compared with the deep learning model with complex structure, only shallow network structure is required. The generalization ability of a single field is relatively strong, with good adaptability to the field.

(3) Comparison of different document lengths

We also compare the influence of different document lengths on classification metrics. The training data of the title, abstract and full-text in the two disciplines are all 400, including 200 positive samples and 200 negative samples. The parameters of causality, loss function and optimization function are controlled unchanged. After iterative training, the performances on CS and Med are shown in table 5. It can be found that, on CS, the best per-

**Table 4**
Ablation experiment: whether causality representation

| Subject → | CS | | | Med | | |
|---|---|---|---|---|---|---|
| Ablation Experiment↓/Metric→ | P | R | F1 | P | R | F1 |
| With Causality Representation | 0.8811 | 0.9056 | 0.8932 | 0.8152 | 0.8721 | 0.8427 |
| Without Causality Representation | 0.8481 | 0.7791 | 0.8121 | 0.7368 | 0.7778 | 0.7568 |

**Table 5**
Metric on different document lengths

| Subject→ | CS | | | Med | | |
|---|---|---|---|---|---|---|
| DifferentLength↓/Metric→ | P | R | F1 | P | R | F1 |
| Title | 0.8750 | 0.8167 | 0.8449 | 0.7857 | 0.7333 | 0.7586 |
| Abstract | 0.7778 | 0.8167 | 0.7967 | 0.7966 | 0.7833 | 0.7899 |
| FullText | 0.9322 | 0.9167 | 0.9244 | 0.9516 | 0.9833 | 0.9672 |

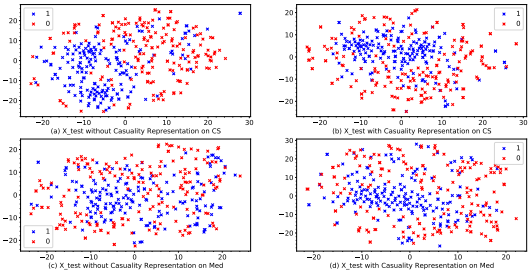


**Figure 7:** Ablation experiment for causality representation.The labels of test data after causality representation have better separability and clearer boundaries, especially CS, while the boundaries between labels without causal representation are fuzzy.

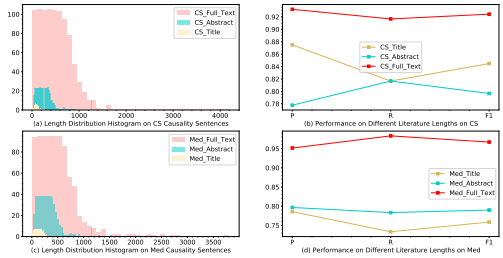


**Figure 8:** Relationship between document length distribution and classification performance. The longer the length, the better the classification performance. For example, CS, the full-text causal sentence has a longer distribution, with P,R and F1 exceeding 0.9.

formance is in the full text, followed by the title. On Med, full text is the best, followed by abstract. The average precision of the two disciplines is 0.9420.

The relationship between length distribution and performance of causal sentences is shown in Fig. 8. It can be seen from the distribution that the precision of different document lengths is positively correlated with the length distribution of documents. The longer the overall distribution of documents tends to be, the higher the accuracy of documents will be. One possible reason is that longer documents capture more information in stage 2, thus releasing effective energy in classification. Therefore, in the future extraction scene, the extraction of full text content should become the main form, which can not only obtain better precision, but also extract more fine-grained knowledge information from the literature, and provide more effective features for more downstream tasks.

## 5. Conclusion

In this paper, we propose a novel framework for scientific literature causality extraction, 2SCE-4SL, which consists of two stages. In stage 1, causal trigger words and entities are identified and collocate into noisy triplets. In stage 2, causality sentences are represented to learn semantic information, this kind of representation can effectively learn the causal dependency and domain-specific knowledge of paper. We compared open scientific literature data in four aspects of the experiments. Results show that the average precision of the 2SCE-4SL is 0.8146, and the F1 is 0.8308, which is the best in the full text with an average precision of 0.9420. And as the training data grows, so does the precision, which makes future optimization possible.

To the best of our knowledge, there are few studies on causality extraction in scientific papers, although this is a direction with great potential, we hope this work can provide enlightenment for relevant research. The principal theoretical implication of this study is that promotes scientific literature mining, especially scientific papers,

to make progress towards deeper natural language inference(NLI) and natural language understanding(NLU), which provides more possibilities for many downstream tasks driven by knowledge and provides a basic guarantee for promoting multidisciplinary knowledge discovery. The principal practical implementation of this study is that 2SCE-4SL only needs a small amount of annotation data to achieve good performance, which provides more possibilities for future performance optimization. In addition, the use of classification framework for causality extraction can reduce the labeling time and improve the efficiency.

Despite these promising results, questions remain. Firstly, 2SCE-4SL extracts only explicit causality and does not include implicit causality. Due to the limitation of causal trigger words, those implicit causalities (i.e. without trigger words but expressing causality) are difficult to be identified. Secondly, the second stage of 2SCE-4SL is represented in sentences rather than paragraphs, which means that those long-distance causality are difficult to identify. Finally, because the framework involves many processes, it remaines to be improved compared with end2end methods, and tag generation requires the intervention of professionals. In the future, we will solve these questions and applications of more downstream tasks.

## Acknowledgments

This research is funded by the National Social Science Foundation of China(Nos.21BTQ071).

## References

[1] L. Bornmann, R. Mutz, Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, Journal of the Association for Information Science and Technology 66 (2015) 2215–2222.

[2] L. Hong, J. Lin, S. Li, F. Wan, H. Yang, T. Jiang, D. Zhao, J. Zeng, A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories, Nature Machine Intelligence 2 (2020) 347–355.

[3] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature, Nature 571 (2019) 95–98.

[4] X. Sun, K. Ding, Identifying and tracking scientific and technological knowledge memes from citation networks of publications and patents, Scientometrics 116 (2018) 1735–1748.

[5] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, A. Zhang, A survey on causal inference, ACM Transactions on Knowledge Discovery from Data (TKDD) 15 (2021) 1 – 46.

[6] S. Bonner, F. Vasile, Causal embeddings for recommendation, in: Proceedings of the 12th ACM conference on recommender systems, 2018, pp. 104–112.

[7] Z. Li, H. Hu, H. Wang, L. Cai, H. Zhang, K. Zhang, Why does the president tweet this? discovering reasons and contexts for politicians' tweets from news articles, Information Processing & Management 59 (2022) 102892.

[8] R. Moraffah, M. Karami, R. Guo, A. Raglin, H. Liu, Causal interpretability for machine learning-problems, methods and evaluation, ACM SIGKDD Explorations Newsletter 22 (2020) 18–33.

[9] J. Liu, Z. Shen, P. Cui, L. Zhou, K. Kuang, B. Li, Y. Lin, Stable adversarial learning under distributional shifts, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 8662–8670.

[10] Z. Li, Q. Li, X. Zou, J. Ren, Causality extraction based on self-attentive bilstm-crf with transferred embeddings, Neurocomputing 423 (2021) 207–219.

[11] T. Dasgupta, R. Saha, L. Dey, A. Naskar, Automatic extraction of causal relations from text using linguistically informed deep neural networks, in: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, 2018, pp. 306–316.

[12] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, et al., Causal inference in natural language processing: Estimation, prediction, interpretation and beyond, arXiv preprint arXiv:2109.00725 (2021).

[13] X. Ding, Z. Li, T. Liu, K. Liao, Elg: an event logic graph, arXiv preprint arXiv:1907.08015 (2019).

[14] J. Xu, W. Zuo, S. Liang, X. Zuo, A review of dataset and labeling methods for causality extraction, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020.

[15] K. Radinsky, S. Davidovich, S. Markovitch, Learning causality for news events prediction, in: Proceedings of the 21st international conference on World Wide Web, 2012, pp. 909–918.

[16] S. Heindorf, Y. Scholten, H. Wachsmuth, A.-C. Ngonga Ngomo, M. Potthast, Causenet: Towards a causality graph extracted from the web, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 3023–3030.

[17] H. Kayesh, M. S. Islam, J. Wang, Event causality detection in tweets by context word extension and neural networks, 2019 20th International Confer-

ence on Parallel and Distributed Computing, Applications and Technologies (PDCAT) (2019) 352–357.
[18] S. Zhao, Q. Wang, S. Massung, B. Qin, T. Liu, B. Wang, C. Zhai, Constructing and embedding abstract event causality networks from text snippets, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017, pp. 335–344.
[19] D.-S. Chang, K.-S. Choi, Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities, Information processing & management 42 (2006) 662–678.
[20] D.-g. Lee, H. Shin, Disease causality extraction based on lexical semantics and document-clause frequency from biomedical literature, BMC medical informatics and decision making 17 (2017) 1–9.
[21] N. An, Y. Xiao, J. Yuan, J. Yang, G. Alterovitz, Extracting causal relations from the literature with word vector mapping, Computers in biology and medicine 115 (2019) 103524.
[22] D. Garcia, et al., Coatis, an nlp system to locate expressions of actions connected by causality links, in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 1997, pp. 347–352.
[23] C. S. Khoo, S. Chan, Y. Niu, Extracting causal knowledge from a medical database using graphical patterns, in: Proceedings of the 38th annual meeting of the association for computational linguistics, 2000, pp. 336–343.
[24] C. Hidey, K. McKeown, Identifying causal relations using parallel wikipedia articles, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1424–1433.
[25] M. Paul, Feature selection as causal inference: Experiments with text classification, in: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), 2017, pp. 163–172.
[26] S. Zhao, T. Liu, S. Zhao, Y. Chen, J.-Y. Nie, Event causality extraction based on connectives analysis, Neurocomputing 173 (2016) 1943–1950.
[27] C. Kruengkrai, K. Torisawa, C. Hashimoto, J. Kloetzer, J.-H. Oh, M. Tanaka, Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.
[28] P. Li, K. Mao, Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts, Expert Systems with Applications 115 (2019) 512–523.
[29] D.-T. Vo, F. Al-Obeidat, E. Bagheri, Extracting temporal and causal relations based on event networks, Information Processing & Management 57 (2020) 102319.
[30] S. Rosenthal, N. Farra, P. Nakov, Semeval-2017 task 4: Sentiment analysis in twitter, in: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), 2017, pp. 502–518.
[31] J. Fu, Z. Liu, W. Liu, W. Zhou, Event causal relation extraction based on cascaded conditional random fields, Pattern Recognition and Artiflcial Intelligence 24 (2011) 567–573.
[32] S. Zhao, M. Jiang, M. Liu, B. Qin, T. Liu, Causaltriad: toward pseudo causal relation discovery and hypotheses generation from medical text data, in: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2018, pp. 184–193.
[33] T. Jiang, T. Zhao, B. Qin, T. Liu, N. Chawla, M. Jiang, Multi-input multi-output sequence labeling for joint extraction of fact and condition tuples from scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.
[34] Y. Zhang, R. Bai, Q. Chen, Y. Zhang, M. Feng, Causal discovery and knowledge linkage in scientific literature: A case study in biomedicine, in: International Conference on Information, Springer, 2022, pp. 319–328.
[35] L. Du, X. Ding, T. Liu, B. Qin, Learning event graph knowledge for abductive reasoning, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5181–5190.
[36] P. Fytas, G. Rizos, L. Specia, What makes a scientific paper be accepted for publication?, in: Proceedings of the First Workshop on Causal Inference and NLP, 2021, pp. 44–60.
[37] P. Wolff, G. Song, Models of causation and the semantics of causal verbs, Cognitive psychology 47 (2003) 276–332.
[38] Z. Luo, Y. Sha, K. Q. Zhu, S.-w. Hwang, Z. Wang, Commonsense causal reasoning between short texts, in: Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning, 2016.
[39] M. Neumann, D. King, I. Beltagy, W. Ammar, Scispacy: Fast and robust models for biomedical natural language processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, pp. 319–327.
[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit,

L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[41] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, L. Bottou, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion., Journal of machine learning research 11 (2010).

[42] Z. Guo, Z. Liu, Z. Ling, S. Wang, L. Jin, Y. Li, Text classification by contrastive learning and cross-lingual data augmentation for alzheimer's disease detection, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6161–6171.

[43] K. Lo, L. L. Wang, M. Neumann, R. Kinney, D. S. Weld, S2orc: The semantic scholar open research corpus, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4969–4983.

[44] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3615–3620.