information

Article

Complex Causal Extraction of Fusion of Entity Location Sensing and Graph Attention Networks

Yang Chen, Weibing Wan *, Jimi Hu, Yuxuan Wang and Bo Huang

School of Electrical and Electronic Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; m020120105@sues.edu.cn (Y.C.); m325121603@sues.edu.cn (J.H.); m320121321@sues.edu.cn (Y.W.); huangbosues@sues.edu.cn (B.H.)
* Correspondence: wbwan@sues.edu.cn

Abstract: At present, there is no uniform definition of annotation schemes for causal extraction, and existing methods are limited by the dependence of relations on long spans, which makes complex sentences such as multi-causal relations and nested causal relations difficult to extract. To solve these problems, a head-to-tail entity annotation method is proposed, which can express the complete semantics of complex causal relations and clearly describe the boundaries of entities. Based on this, a causal model, RPA-GCN (relation position and attention-graph convolutional networks), is constructed, incorporating GAT (graph attention network) and entity location perception. The attention layer is combined with a dependency tree to enhance the model's ability to perceive relational features, and a bi-directional graph convolutional network is constructed to further capture the deep interaction information between entities and relationships. Finally, the classifier iteratively predicts the relationship of each word pair in the sentence and analyzes all causal pairs in the sentence by a scoring function. Experiments on SemEval 2010 task 8 and the Altlex dataset show that our proposed method has significant advantages in solving complex causal extraction compared to state-of-the-art methods.

Keywords: entity location awareness; long span dependence; GAT; GCN; complex causal extraction

Citation: Chen, Y.; Wan, W.; Hu, J.; Wang, Y.; Huang, B. Complex Causal Extraction of Fusion of Entity Location Sensing and Graph Attention Networks. *Information* **2022**, *13*, 364. https://doi.org/10.3390/info13080364

Academic Editor: Haridimos Kondylakis

Received: 14 June 2022
Accepted: 28 July 2022
Published: 31 July 2022

1. Introduction

Decision-making is an important act of analyzing and reasoning about a problem, and the correctness or incorrectness of a decision often determines the success or failure of an activity. For some simple decisions, people can use knowledge and experience accumulated in the past to decide by intuitive and qualitative analysis. With the development of science and technology, however, the decision-making environment has become more open, and the uncertainties in the decision-making process have become increasingly numerous; furthermore, the data available for decision-making have become increasingly large-scale and diverse, and they contain a very large amount of information and clues relating to cause–effect relationships that need to be analyzed.For example, the left-hand side of Figure 1 shows three real posts from Twitter. By considering these common-sense causes and effects, we can extract the reasoning content: "because of rain, we catch a cold"; rain is the cause of a cold, and a cold is the result of rain.

In cause-and-effect relationships, the occurrence of a cause triggers the occurrence of an effect. This forms the basis of inferences [1] and decisions [2] and plays a pivotal role in event analysis [3], logic models [4], decision judgments [5], and other application scenarios. The example in Figure 1 shows the importance of causal extraction. As noted, this gives us the following results: getting caught in the rain leads to colds; many types of viruses can cause a common cold. If we introduce expert experience into common-sense causality, using causal analysis of clinical observation data gives us: "consume warm foods and warm drinks" to treat "rhinovirus"; and "influenza virus" is treated with "antiviral drugs".

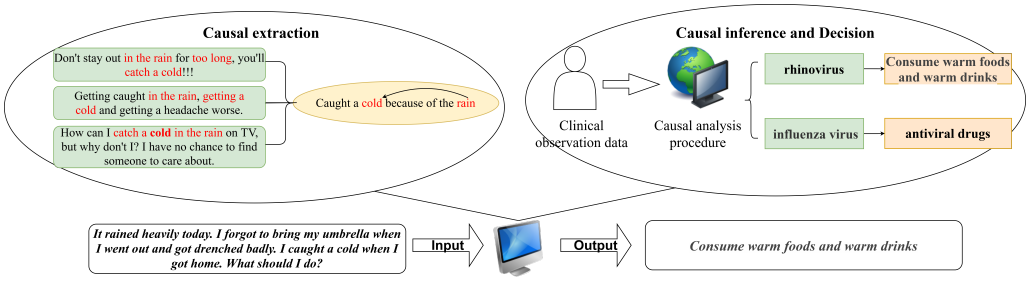

Figure 1. Application of causality extraction in decision-making and question answering.

On the premise of comprehensive use of the causal conclusions obtained above, if the new case is described as: "It rained heavily today. I forgot to bring my umbrella when I went out and got drenched badly. I caught a cold when I got home. What should I do?" Using the knowledge about cold types, we can infer that the patient has rhinovirus and that the optimal solution should be to consume warm foods and warm drinks. This is an important semantic relationship that reflects a strong correlation between two entities before and after from cause to effect. Identifying and extracting causality from unstructured sentences—and the precise determination of cause and effect in relation to entity boundaries—is of great significance for high-level applications such as precise causal inference and event-development inference chains. As such, many researchers are devoted to the study of causality extraction.

In recent years, there has been great progress in research examining the extraction of causal relationships. However, most previous work has focused on intra-sentence relationships and simple entity–pair relationships; there are still two main challenges that remain. First, as in sentence (a) in Figure 2, in long-sentence relationship extraction, the entity pairs corresponding to the relationship usually span multiple words, and the existing methods are limited by the dependence of the relationship on this long span; accurately unifying the entity information and the feature representation of the long-span semantic information still needs to be explored. Second, as in sentences (b)–(d) in Figure 2, when the same causal entity appears in multiple relationship triads (one cause and multiple effects or chain-causal cases), the classifier is prone to confusion; without the support of a large amount of training data, it is difficult for the classifier to discern the relationships in which the entities are involved, and this can greatly affect the performance of the model.

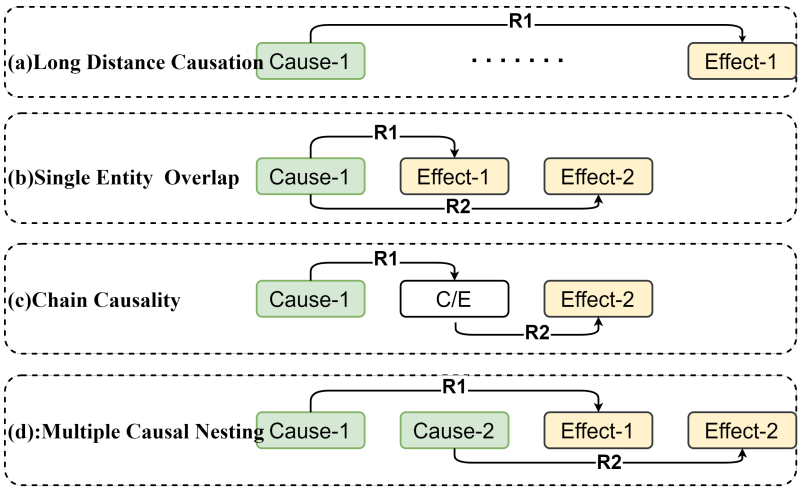

Figure 2. Schematic diagram of complex cause–effect relationships.

It is still difficult to accurately identify complex causal relationships using semantic information. To address the above problems, herein, we propose a new causality-extraction method that incorporates an entity-location-aware graph-attention (GAT) mechanism to reduce redundant content in graph-dependency trees and enhance the association between

long-span entities. We conducted a large number of experiments to test this approach, and the results showed its effectiveness. The main contributions of this work are as follows:

1. To address the problems of overlapping multiple relationship groups and long-span dependencies in causal-relationship extraction, a novel head-and-tail annotation scheme for relationship entities is proposed. This contains head and tail nodes for causal entities and relationship words, dividing the triads in the text into multiple simple small sets according to relationship categories and reducing the complexity of subsequent entity recognition. Entities of arbitrary span can be detected using these head and tail pointers, which capture information-boundary data and define scoring functions. All possible mentions of a causal entity in a sentence can be detected iteratively, fully integrating the interaction information between entities and relationships. This provides notable advantages for solving complex causal and long-span causal problems without complex feature engineering.
2. We propose a GAT mechanism incorporating entity-location perception, in which entity-location perception strategies provide constraint guidance for graph-dependency semantics to learn relationships between long-span nodes and reduce redundant interference. This enables better capture of long-range dependencies between entities and strengthens the dependency-association features between causal pairs.
3. We build a bidirectional graph convolutional network (GCN) to perform deep mining of the implicit relationship features between each word pair outputted from the attention layer. This improves the directionality of the subject and object in relationship extraction, and it allows iterative prediction of the relationship between each pair of words in a sentence using a classifier; the functions are then scored to analyze all causal pairs in a sentence. Experiments were conducted on a sentence-level explicit-relationship-extraction corpus. The experimental results show that the proposed method obtains the optimal F1 value when compared with the state-of-the-art model, and it effectively improves the extraction accuracy for complex cause–effect and long-span sentences.

2. Related Work

This section presents a survey of the previous research related to causal-relationship extraction and briefly describes causal-sequence labeling methods and relationship-extraction techniques.

2.1. Causal-Sequence Labeling Method

Dasgupta et al. [6] labeled sentences based on four tags—cause (C), effect (E), causation (CC), and none (None)—dividing a sentence into multiple clauses, with each relationship considered as a separate sentence. Fu et al. [7] used tags to indicate the semantic role of causation of events ("cause" (C), "result" (E), or "other" (N)) and tagging labels to indicate the boundaries of event causality (an event is causally external (O), initiating causality (B), or continuing causality (I)) in two steps to assign labels to each event in the sequence and to identify the degree of causality by boundary labels. However, this annotation scheme introduces a large number of new causal labels, which leads to an uneven distribution of semantic annotations and creates difficulties in feature training. Zhao et al. [6] applied the Cartesian product of entity-mention tokens and relationship-type tokens and then assigned a unique token to encode the entity mention and relationship type for each word. Although these studies achieved good improvements in causal-relationship extraction, all of the approaches treated relationships as discrete labels to be assigned to entity pairs, making relationship classification a difficult machine-learning problem: assuming that a complex relationship extraction contains m entities and each entity pair may contain relationships between them, there are A_m^2 possible combinations of entity pairs, and these are prone to a large number of negative examples.

To address these problems, Wei et al. [8] proposed a new cascaded binary-tagging framework for triple extraction of relationships that identifies the tail entities under each

specified relationship using the premise of identifying the head entities first. This can solve the relationship-overlap problem to a certain extent, and it achieves a great enhancement for the extraction of overlapping relationships. The method also partly alleviates the problem of there being too many combinations of entity pairs, but there is still much room for improvement.

In the approach described in this paper, we determine entity boundaries by encoding the start and end positions of causal-entity pairs to improve entity-position-aware features. This effectively alleviates the long-distance-dependence problem of relationship extraction. By identifying all possible causal-entity pairs mentioned through the joint extraction of entities and relationships, the primary features of the entities are extracted as their own representations for matching. The relationship-extraction problem is then transformed into an identification problem of three elements: relationships, head entities, and tail entities.

2.2. Relationship Extraction Technology

Traditional causality extraction is divided into two types of approaches, rule-based and machine-learning based. Garcia [9] et al. defined a lexical syntax of 23 explicit causal verbs, and the rules explore the causal relationships present in the context by finding field operations. One of them, Kira [10], proposes an authoritative algorithm for generating causal pairs from news articles. This rule-based approach achieves a higher degree of automation than the two pattern-based systems mentioned above, thanks to the use of generalized rules <pattern, constraint, priority>. The system achieves an accuracy of 70.4% for news article titles over a period of 150 years. However, the system achieves a poor recall metric of 10.0% because the rules only cover obvious causal cases.

Statistical-based machine learning reduces manual predefinition relative to rule-based approaches. Zhao et al. [11] avoided the overfitting problem of hidden plain Bayesian models by handling partial interactions between text features and introduced a new causal linkage feature that classifies linking words into different categories. Kim et al. [12] combined probabilistic topic models with a time-series causal analysis to mine causal relationships. Lin et al. [13] constructed a classifier to mine relationships with four rules: generative rules, dependency rules, word pairs, and context. The model has a high recall but low precision. This is because causal relations are the most dominant relations in the training set, and this class of models tends to label uncertain relations as causal instances. These traditional methods are limited to extracting the semantic features of sentences and ignore the dependent association features in sentences.

With the maturity of deep learning techniques, training models map words and features into low-dimensional dense vectors to effectively alleviate the feature sparsity problem. Wang et al. [14] proposed a multisegmental attention-based CNN model to capture information specific to entity relations, and an attention pooling layer is used to capture the most useful convolutional features of the CNN; however, this treatment ignores the long-range dependencies among causal relations. Mao [15] et al. accurately represent causal keywords and cue phrases by word filtering techniques, output a similarity score for each word filter, and combine clustering techniques to downscale the features to improve the convolutional performance of CNN models. However, the pooling layer misses some valuable information, cannot bridge the correlation between the local vector of text and the whole, and ignores the long-range dependency between causal relations. Xu [16] et al. used the shortest dependency path (SDP-LSTM) to extract heterogeneous information during relation classification to learn higher-level semantic and syntactic representations. Li et al. Zhao [17] used textual causal connectives to identify causal pairs and extracted specific causal events from the identified sentences, obtaining high-quality and readable causal pairs. Li [18] et al. combined Bi-LSTM with a multi-headed self-attention mechanism to direct attention to long-range dependencies between words. Graph-based models in locally continuous word sequences, such as graph convolutional networks (GCNs) and graph attention networks (GATs) to model a set of points (nodes) and their relations (edges), have also received attention from researchers. Zhang et al. [19] proposed a dependency

tree-based GCN model to extract relations. The pruning strategy is applied to graphical convolutional neural networks, which can remove irrelevant content without ignoring key information and thus improve the robustness of the model. The literature [20] introduced syntactic dependency trees in the graph attention network layer to enhance causal-semantic associations but cannot represent accurate sequence labeling for complex relations.

Despite the success of these joint approaches, they all assume that there are no overlapping triads in a sentence and the long-span problem of relational pairs is not better addressed, so the dependency feature between causal pairs is crucial to facilitate inference about relational triads. Dai [21] directly labels entities and relational tags based on query word positions and identifies entities in other positions that are related to the former. Dixit [22] constructs a span-level graph for joint detection of overlapping entities and relations. However, these models cannot achieve satisfactory performance when the overlap is relatively complex. Zhang et al. [23] combined entity location with perceptual attention to encode semantic information and global location of entities. The ablation study showed that this location-aware mechanism was effective and improved the F1 value by 3.9%. Inspired by this, we propose a new causality extraction method that fuses entity location-aware graph attention mechanisms to reduce redundant content from graph dependency trees and thus enhance the dependent association features between long-span entities.

3. Methods

In this work, the essence of causal-relationship extraction is the automatic tagging of causal subject–object words in the corpus. In this section, we describe the detailed principles of our approach. The causal word tagging method is introduced in Section 3.1, and the causal relationship extraction model (relationship position and attention (RPA-GCN) is introduced in Section 3.2). As shown in Figure 3, for the input sentences, RPA-GCN is used for feature encoding and relationship prediction. In the subject-tagging stage, the decoding of relationship classification is input to the subject-tagging network as a precondition. In the object-entity labeling stage, the relationship and subject-entity decoded in the first two stages will be input as preconditions to further limit and enrich the feature representation of this layer. The classification function predicts the relationship between each pair of words in the sentence, and the relationship prediction at this time is converted into multiple binary-classification tasks, which greatly reduces the complexity of triples. Through the connection and progression of these three stages, the semantic representation of the relationships and entities in the model network is gradually enriched, and the complex causal relationship existing in the sentence is well discriminated.

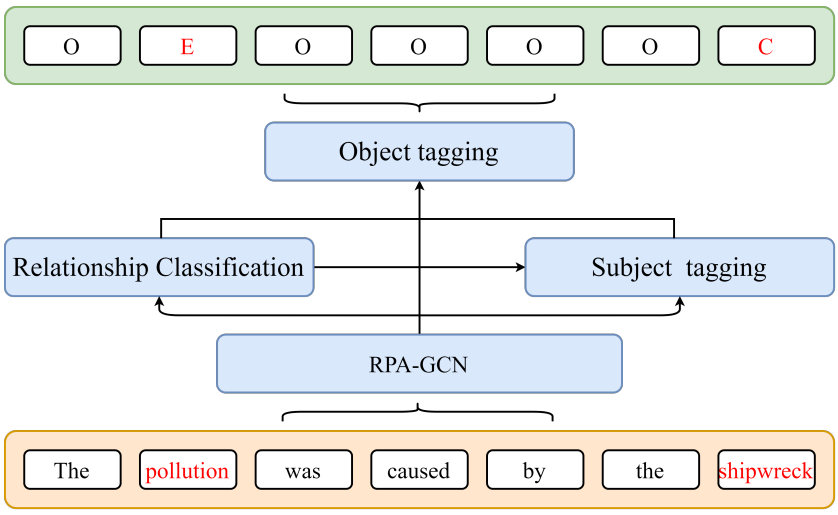

Figure 3. Research steps.

3.1. Cause–Effect Sequence Labeling Method

In our method, due to the limited number of applicable causal corpora available at present, we avoid using complex tokens that might cause the reduction in causal semantic purity, and we use "C" (cause), "E" (effect), "CE" (cause and effect), and "O" (other) as causal tags. The compound tag CE indicates that the entity is the result of a first causal pair and the cause of a second causal pair, which intelligently addresses the embedded causality. To enhance the isolation of entity boundaries and relational edges and to improve the pruning effect on the dependency tree, the causal extraction is decomposed into two interrelated sub-tasks: the identification of starting and ending positions of causal-entity pairs. In Figure 4, the position corresponding to "high" in the cause entity "high inflation" is labeled as "C_h", which indicates the cause head, and the position corresponding to "inflation" in the end marker sequence is labeled as "C_t", which indicates the cause tail. The single causal-entity word is marked as "C/E", and the unrelated word vector is marked as "O". Therefore, the process of causal-pair extraction is to determine the boundaries of causal entities and define the relative-position code by identifying the start and end positions of the entity. The relative position of "Recessions" to "downward price rigidity" is −11.

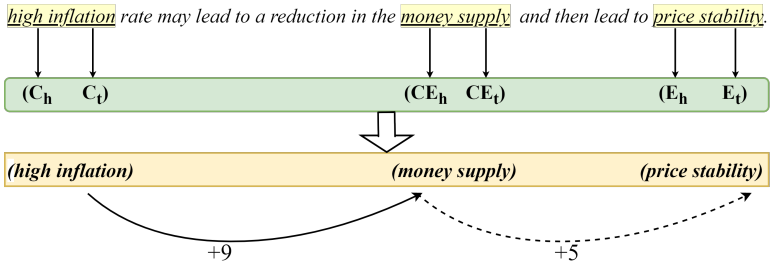

Figure 4. Chain causal sentence pattern.

To solve the problem of causal chaining or nesting in sentences and to alleviate the problem of inadequate long-span-dependent features between causal entities, we determine entity boundaries by encoding the beginning and ending positions of causal-entity pairs to improve relative entity-position-aware features. The joint extraction of entities and relationships identifies all possible causal-pair-entity mentions, extracts the primary features of the entities as their own representations for matching, and transforms the causal-extraction task into a sequential annotation of the relationship group (C_h, C_t, E_h, and E_t).

For each start position in the annotation result, it is known from the common sense of the text that the end position must be after the start position, so annotation results before the start position can be ignored. Since the span of any entity in a sentence is continuous, the accuracy of the entity can be guaranteed as long as the start position and end position of the entity can be correctly identified. In Section 3.2.3, we will describe how the head and tail pointers match the cause and effect in the sentence parsed by the scoring function. Compared with the method of a previous study [6] , which simply cut the sentences into two parts based on the causal conjunction, the annotation scheme in this method effectively solves the problems of overly loose causal boundaries and uneven distribution of tokens.

3.2. Causality-Extraction Model (RPA-GCN)

The RPA-GCN model, which fuses entity-position perception and GAT for capturing sentence-level-dependent features, is proposed in this section, as shown in Figure 5. The model goes through several feature-extraction processes as follows. (1) Sentences are passed through the word-embedding layer to obtain the initial feature representation of the sequence, and the feature representation obtained from the word-embedding layer is input to the bidirectional long short-term memory (Bi-LSTM) layer. The distance of each word vector relative to the entity is calculated as its relative position feature, and this is also input to the Bi-LSTM layer. The LSTMs of the front- and back-directional sequences are then

generated into long-span encoding vectors. (2) The GAT mechanism learns the potential features of the coding dependency tree, and the Bi-LSTM layer outputs relative-position information, aggregates the effective information in the dependency tree, and enhances the associations between related entities. (3) The bidirectional GCN further learns the deeper dependency interaction features of the semantics, and a pooling operation reduces the redundancy of the feature information and parses all the causes and effects in the sentence.

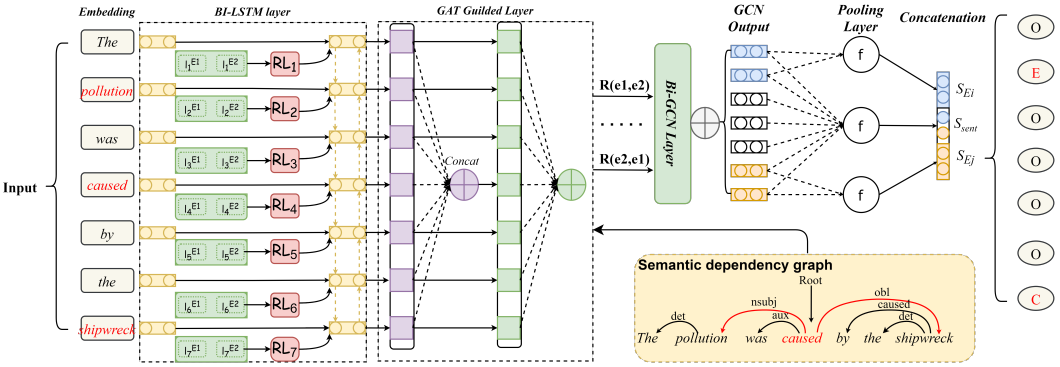

Figure 5. Framework diagram of causal extraction model (RPA-GCN).

3.2.1. Bi-LSTM Layer (Fused Entity Location Information)

As shown in Figure 6,to alleviate the inadequate extraction of long-span-dependent features for causality extraction, entity location-aware information is introduced in the Bi-LSTM layer. l is the number of words in an entity pair, and E_t is the vector of the t entity pair. Let the first relational entity in the sentence be represented by E1 and the second by E2. Calculate the distance l_t^{E1} of the current word t with respect to the first entity and the distance l_t^{E2} with respect to the second entity.

$$E_t = \frac{1}{l}\sum_{c=1}^{l} h_c \tag{1}$$

$$l_t^{E1} = l_t - l_{E1} + L \tag{2}$$

$$l_t^{E2} = l_t - l_{E2} + L \tag{3}$$

$$Rl_t = \left[w_t, l_t^{E1}, l_t^{E2}\right] \tag{4}$$

where w_i is the word vector of the current word i, l_{E1} and l_{E2} are the position codes of entities E1 and E2, respectively. L is the maximum length of the sequence in the current training batch.

To obtain semantic contextual continuous sequence features, the input is encoded using the Bi-LSTM. The LSTM uses a gating device to achieve long-term memory, which alleviates long sequence forgetting information, and it is also able to capture sequence information. The hidden state representation in the Bi-LSTM is shown in Equations (5)–(8).

$$\overrightarrow{h_i} = \overrightarrow{LSTM}\left(g_i, \overrightarrow{h_{i-1}}\right) \tag{5}$$

$$\overleftarrow{h_i} = \overleftarrow{LSTM}\left(g_i, \overleftarrow{h_{i-1}}\right) \tag{6}$$

$$h_i = \left[\overrightarrow{h_i}; \overleftarrow{h_i}\right] \tag{7}$$

$$Wh_i = h_i * W_c + b_c \tag{8}$$

where g_i is the result of stitching using different convolution kernels. $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ denote the forward and backward hidden states of h_i in the Bi-LSTM, respectively, and the extracted features are spliced as h_i, and the feature is represented by Wh_i after h_i passes through the linear layer and as the output of the Bi-LSTM layer.

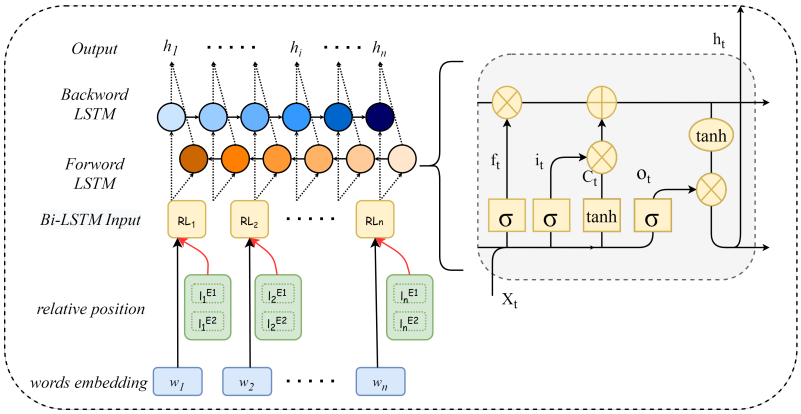

Figure 6. Bi-LSTM network structure.

3.2.2. GAT Layer

A GAT network is a new type of neural-network architecture based on graph-structured data that is able to engage the features of its neighbors by stacking layer nodes. This means that the attention is more focused on the relationships between the words to be extracted, further strengthening the semantic features, incorporating the complete factual information sentence by sentence to update the graph network, and more comprehensively exploring the relationships between entities. We parse the dependency information of the sentences to obtain the adjacency matrix that stores their semantic information. This is input to the GAT layer along with the output features of the Bi-LSTM layer, and this incorporates entity-location awareness for further feature extraction.

A semantic dependency pattern is a linguistic structure that represents the relationship between entities in a sentence. In the Stanford Parser's analysis of syntactic dependency [24], the constraint information of lexical dependency reduces the interference of redundant words. Take the sentence "The pollution was caused by the shipwreck" as an example. Here, "caused" is the root node, and the causal subject and object point to "pollution" and "shipwreck". The semantic dependency graph is stored in the form of an adjacency matrix, which is a directed acyclic graph. The dependency tree and adjacency matrix of sentence generation are shown in Figure 7.

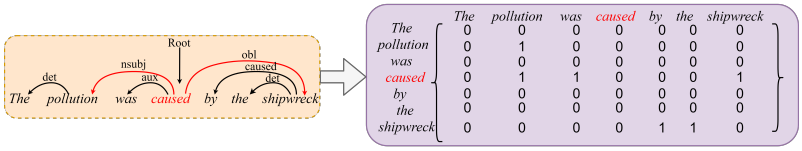

Figure 7. Dependency tree example of causal relational sentence pattern.

In the GAT layer, the Bi-LSTM output coding incorporating the entity-location-aware mechanism performs constraint guidance on the dependency semantics to learn the relationships between long-span nodes and to reduce redundant interference. As shown in Figure 8, the nodes of the graph correspond to each word in the sentence, the node features are the word features extracted from the Bi-LSTM layer, and the edges of the graph correspond to the edges of the semantic dependency graph, i.e., the corresponding semantic

relationships in the adjacency matrix. The Bi-LSTM layer outputs fused entity-location-aware feature vectors, and these are concatenated and input to a single-layer feedforward neural network a, and e_{ij} is obtained by a nonlinear activation function:

$$e_{ij} = a(\mathbf{W}h_i, \mathbf{W}h_j) = LeakyReLU\left(W_a^T\left[\mathbf{W}h_i \| \mathbf{W}h_j\right]\right) \tag{9}$$

where the linear mapping of the shared parameter W adds dimension to the features of the vertices (this is a common feature-augmentation method), the operator "$\|$" splices the transformed features of vertices i and j, and $a()$ concatenates the spliced features. The latter high-dimensional features are mapped to the real number e_{ij}, which denotes the importance of word j to word i.

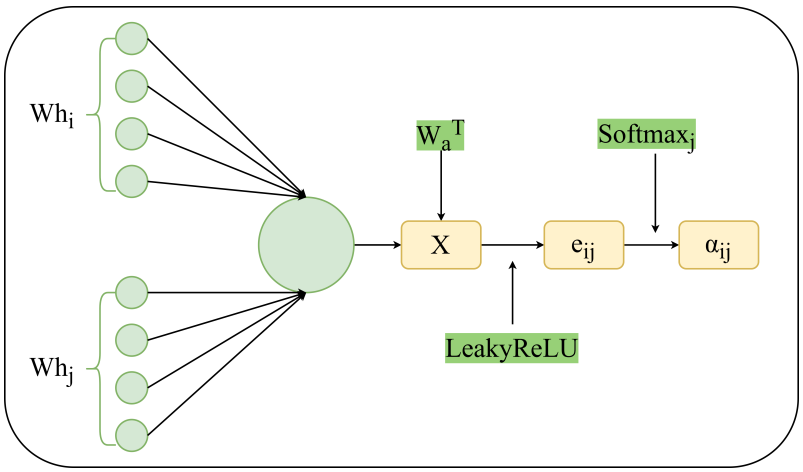

Figure 8. Attention coefficients.

The computation of GAT is performed by mining the dependencies between words through the adjacency matrix generated by semantic dependency analysis. The normalized attention coefficient α_{ij} is calculated as:

$$\alpha_{ij} = softmax_j(e_{ij}) = \frac{exp(e_{ij})}{\sum\limits_{k \in D_i} exp(e_{ik})} = \frac{exp(LeakyReLU\left(W_a^T\left[\mathbf{W}h_i' \| \mathbf{W}h_j'\right]\right))}{\sum\limits_{k \in D_i} exp(LeakyReLU(W_a^T\left[\mathbf{W}h_i' \| \mathbf{W}h_k'\right]))} \tag{10}$$

where the set of nodes adjacent to word i in the semantic dependency graph is D_i, and the softmax function is introduced to normalize all the nodes j adjacent to i. All the neighboring word features j of word i in the dependency graph with corresponding weight coefficients α_{ij} are weighted and summed to obtain attentional features h_i' using a nonlinear layer:

$$h_i' = \sigma\left(\sum_{j \in D_i} \alpha_{ij} \mathbf{W}h_j'\right) \tag{11}$$

3.2.3. GCN Layer

The GCN used in this method is a variant of the undirected GCN proposed in a previous report [25]. This aims to capture high-level neighborhood information and learn the representations of nodes in convolutional operations on the graph. To further extract the interaction information between entities and relationships, the GCN performs deep mining on the implied relationship features between each word pair output from the GAT layer; however, an undirected graph structure cannot represent the directionality of the subject and object in the causal relationship. This is improved in this paper by constructing a GAT-guided bidirectional GCN network. This considers the relative position weights between sentence entities to deal with the complex causal-entity-pair distribution problem.

The parameter h_i^l denotes the output features of the ith node in the lth layer of the GCN, and h_i^{l-1} denotes the corresponding input feature. These are calculated as follows:

$$\overrightarrow{h}_i^l = \sigma\left(\sum_{j=1}^{n} A_{ij}\left(\overrightarrow{W}_h^l q_j^{l-1} + \overrightarrow{W}_{rel}^l R_{ij}\right) + b^l\right) \tag{12}$$

$$\overleftarrow{h}_i^l = \sigma\left(\sum_{j=1}^{n} A_{ij}\left(\overleftarrow{W}_h^l q_j^{l-1} + \overleftarrow{W}_{rel}^l R_{ji}\right) + b^l\right) \tag{13}$$

$$h_i^l = \left[\overrightarrow{h_i}; \overleftarrow{h_i}\right] \tag{14}$$

where W^l is the weight matrix, b^l is the bias vector, σ is the sigmoid function, and R_{ij} is the relational weight representation of the self-attentive layer nodes for (i, j).

$$S_{sent} = f\left(S^{(L)}\right) = f(GCN(S^{(0)})) \tag{15}$$

$$S_{Ei} = f(O_{Ei}) \tag{16}$$

where $S^{(L)}$ represents the hidden representation obtained by the GCN in layer L, f represents the maximum-pooling function, which maps n output vectors to a sentence vector. The parameter S_{Ei} is the entity-classification representation, the sentence representation S_{sent} is spliced with the relational-entity head-end representation S_{Ei}, and the final sentence-feature representation is obtained by the feedforward neural network (FFNN):

$$S(i, j) = FFNN\left(\left[S_{sent}; S_{Ei}; S_{Ej}\right]\right) \tag{17}$$

For each relationship, the weight matrix $\mathbf{W}_r^1\mathbf{W}_r^2\mathbf{W}_r^3$ is learned, and the propensity score between each two words in the sentence belonging to the relationship r is calculated as:

$$S(i, j|RL) = \sigma\left(\mathbf{W}_{RL}^3 ReLU(\mathbf{W}_{RL}^1 S_i \oplus \mathbf{W}_{RL}^2 S_j)\right) \tag{18}$$

For each relationship RL, we learn the weight matrices W_{RL}^1, W_{RL}^2, and W_{RL}^3 and compute the score $S(i, j|RL)$ of the relationship trend, with $W_{RL}^1 S_i \oplus W_{RL}^2 S_j$ representing the posterior probability modeling of the relationship class for i as subject and j as object. Here, $S(i, j|RL)$ denotes the relational propensity score of word pairs (i, j) under the relationship RL. Because there is no bidirectionality in the relationship between causal pairs, $S(i, j|RL)$ is different from $S(j, i|RL)$. Saha and Kumar [26] classified message tone using n-grams to select an appropriate emoji. In the present approach, the relationship between each pair of words in a sentence is predicted using the sigmoid function, and this is converted into multiple binary-classification tasks. Then, all possible relationships are given a larger prediction probability for a sentence with multiple relationships for the same word pair, thus alleviating the relationship overlap problem.

The parameter S_{final} is used as a scoring function to analyze cause and effect in a sentence:

$$S_{\text{final}} = S(\mathrm{C_h}, \mathrm{C_t}) + S(\mathrm{E_h}, \mathrm{E_t}) + S(\mathrm{C_h}, \mathrm{E_h}|RL) + S(\mathrm{C_t}, \mathrm{E_t}|RL), \tag{19}$$

where: $S(\mathrm{C_h}, \mathrm{C_t})$ and $S(\mathrm{E_h}, \mathrm{E_t})$ are the first and last scores of cause and effect, respectively; $S(\mathrm{C_h}, \mathrm{E_h}|RL)$ and $S(\mathrm{C_t}, \mathrm{E_t}|RL)$ match the first and last features of cause and effect as the subject and object representations, respectively; and the constraint $S_{\text{final}} > 0$ is imposed to analyze all relationships for all possible causes and effects. The loss function L is set as a negative log-likelihood function:

$$L = -\sum_{i=1}^{d}\sum_{j=1}^{m} y_{ij}\log(\hat{y}_{ij}) + \lambda\|\theta\|^2, \tag{20}$$

where y is the sample relationship label, $\hat{y}$ is the predicted relationship label, d is the total number of samples in the training set, m is the number of relationship categories, λ is the L2 regularization coefficient, and θ is the model parameter.

4. Experiments and Analysis

4.1. Experimental Data

In the SemEval-2010 task 8 dataset [27], there are 10,717 examples, and nine types of relationships and "other" types, one of which is causality. There are 1368 sentences with cause–effect relationships and 107 sentences without cause–effect relationships. Causes and consequences can be directly derived from the labeled causal entities and the direction of causality, but the small sample size and the unbalanced conditions are the main limitations of this dataset.

The AltLex dataset [28] extracted the text of sentences with causal relationships from the English Wikipedia corpus. There are 4595 causal sentences and 39,645 non-causal sentences, which is a large amount of data. The aim of the present work was to explore the influence of the location relationships of entities on the extraction performance. Therefore, implicit causal sentences were filtered out from this corpus, and 2482 explicit causal sentences were extracted for subsequent experiments.

The dataset for this paper was selected to extend the SemEval-2010 task 8 and AltLex data and to address their shortcomings as follows:

1. *The existence of multiple cause − effect relationships.* As in the example of Figure 9, the relationship labeling of the original corpus sentences is limited to one cause and one effect. This ignores the possible existence of multiple cause − effect relationships in most of the sentences. We expand the candidate cause − effect pairs for the sentences.
2. *Presence of chain causality.* As in the example in Figure 4, a word can simultaneously be a cause or effect in multiple causal pairs; herein, the sentence treatment is considered an extraction of multiple sets of causal pairs. The model is more generalized than that of a previous study [29], which focused on the most basic causes for chained causal sentences.

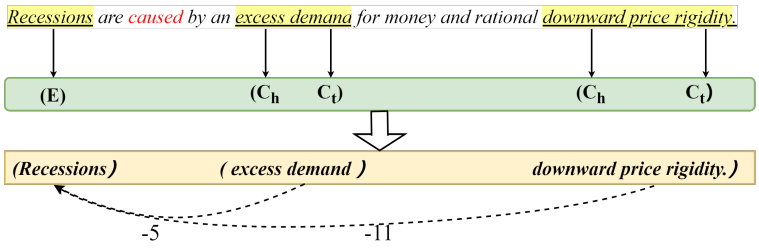

Figure 9. Multi−causal sentence pattern.

Ultimately, 3850 causal sentences and 4769 randomly selected non-causal sentences were obtained to constitute the dataset. The sentences were analyzed for dependent syntax using Stanford CoreNLP [30], and 80% of the results of an intensive audit of automatic annotation were randomly selected as the training set, 10% were used as the validation set, and 10% were used as the test set.

4.2. Experimental Parameter Setting

In the linear layer, the input vector dimension of the model was set to 600, the output vector dimension was 300, and the maximum entity length was 8. In the relative-position-encoding module, the input size of the Bi-LSTM was 360, the hidden size was 300, the number of layers was 1, the dropout was 0.5, the dimension of the position embedding was set to 30, the network weight optimizer was Adam, and the dropout was set to 0.25. The batch size was set to 8, and the learning rate was 1×10^{-4}.

4.3. Evaluation Indicators

The model test was performed at the sentence level for sequence labeling, and the accuracy of this was determined based on the results of the labeling sequences and relationship types. All the words in the sentence were correctly labeled: (1) the starting and ending boundaries of the causal entities were completely correct; (2) the types and directions of the relational entities were correct; (3) the above conditions were also satisfied in complex sentences such as those with chain causation and multiple causation.

We used the standard precision (P_i), recall (R_i), and F1 scores to evaluate the experimental results. These can be calculated using:

$$P_i = \frac{|A_i \cap G_i|}{|A_i|} \tag{21}$$

$$R_i = \frac{|A_i \cap G_i|}{|G_i|} \tag{22}$$

$$F1 = \frac{2P_iR_i}{P_i + R_i} \tag{23}$$

where A_i is the extracted causal triad, G_i is the number of causal triads in the set of all sentences in the dataset, and the predicted relational triad is considered correct when and only when the relationship and the boundaries of the subject and object are correct.

4.4. Baseline Model Comparison

In this paper, we use the same experimental setup and compare it with the mainstream research methods using the SemEval 2010 task 8 dataset so as to validate the performance of the model for multi-relational classification. We compare with the following benchmark models.

LR [23]: A traditional relationship extraction model based on dependency trees combined with lexical information.

SDP-LSTM [27]: a multichannel recurrent neural network with long short-term memory (LSTM) units is used to pick up heterogeneous information along the SDP (shortest dependency path between two entities) and retain the most relevant information.

PA-LSTM (2017) [23]: combines an LSTM sequence model with a form of entity location-aware attention that is more suitable for relationship extraction, employing an entity location-aware attention mechanism on the sequence model. This model does not use dependency trees.

C-GCN (2018) [19]: introduced graph convolutional networks to detect entity relationships and employed pruning strategies based on dependency analysis to maximize the removal of irrelevant content around SDPs.

C-AGGCN (2020) [31]: is a soft pruning strategy based on an attention mechanism that takes the whole dependency tree as input and applies a multi-headed self-attentive mechanism to measure the weights of each node on the dependency tree and incorporates dense connectivity.

As can be seen from Table 1, our model achieves an F1 Score of 85.67 and performs best on the mean F1. The relative F1 of SDP-LSTM [16] improves by 1.5% relative to LR [23], which indicates that the dependency path between entities can effectively assess the weight of attention of relational entities throughout the sequence. The GCN-based models in the literature [19,31] all embed syntactic structures, and obviously, the GCN-based models perform better in the F1 test. RPA-GCN contains contextual information to capture long-range dependencies at the syntactic level, incorporates entity location awareness in computing graph attention weights, weights the neighborhoods of pruned graph-dependent nodes, and enhances the semantic features of relational entity pairs, which is more advantageous compared to other GCN models.

Table 1. SemEval2010 task 8 relational extraction of F1.

Model	F1
LR [23]	82.2
SDP-LSTM [16]	83.7
PA-LSTM [23]	82.7
C-GCN [19]	84.8
C-AGGCN [31]	85.1
RPA-GCN(Model of this paper)	85.67

Based on the causal corpus in Section 3.1, we compare the precision, recall, and F1 values of the GCN-based baseline model, as shown in Figures 10–12.

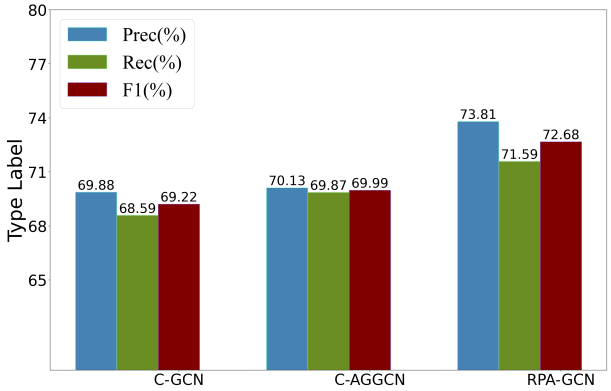

Figure 10. Value of precision, recall, and F1 of relationship type.

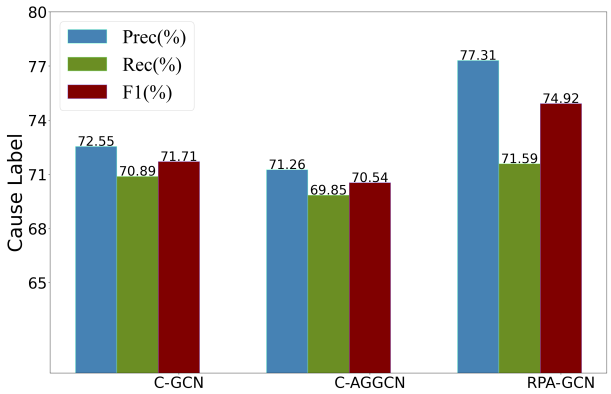

Figure 11. Value of precision, recall, and F1 of cause tags.

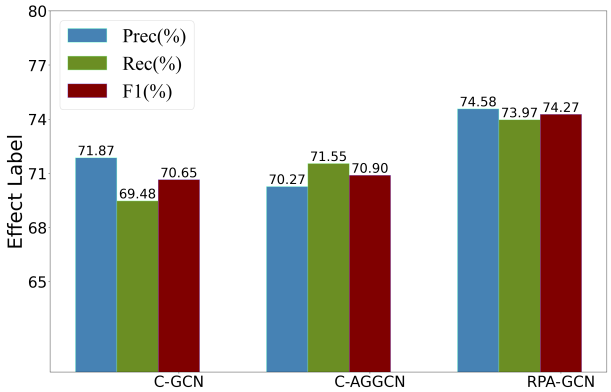

Figure 12. Value of precision, recall, and F1 of effect tags.

As can be seen from Figures 10–12, the precision and recall of RPA-GCN are substantially better than those of other baseline models, with and without relative position encoding, with a difference of four percentage points. The weight-learning strategy of relative encoding information of self-attention models the position of subject and object in the sequence or the global position of entities within the sequence and more clearly portrays entity boundaries in the feature extraction process, making the causal entity features richer.

To verify the stability of the model, we analyzed the effect of the number of iterations on the F1 value, and we tested the variation of the F1 value obtained by the three methods with different numbers of iterations.

Figures 13 and 14 show the training loss and the best F1 score on the validation set for the baseline model [19,31]. The F1 value gradually increases in the first 19 iterations, and then it converges to stabilize. Although there are some fluctuations, these are not large, demonstrating the stability of the model. Further, we analyzed the performance of the baseline model under different training data sizes.

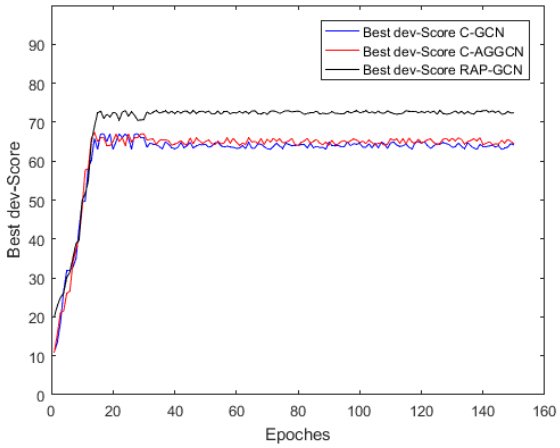

Figure 13. Epoch-F1 (Val).

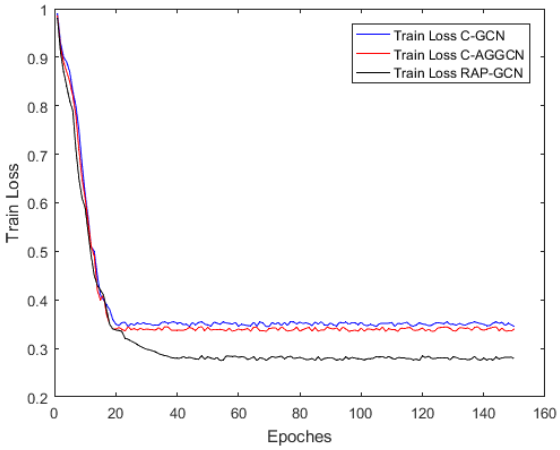

Figure 14. Epoch-loss (train).

We set five training data sizes (20%, 40%, 60%, 80%, and 100%) and evaluated the performance of the C-GCN, C-AGGCN, and RPA-GCN models. As shown in Figure 15, RPA-GCN consistently outperforms the baseline model for the same quantity of training data, and the performance gap between the three models becomes more pronounced as the size of the training dataset is increased above 60%. These results suggest that our models are more effective in terms of using training resources. Finally, we analyzed the model ablation.

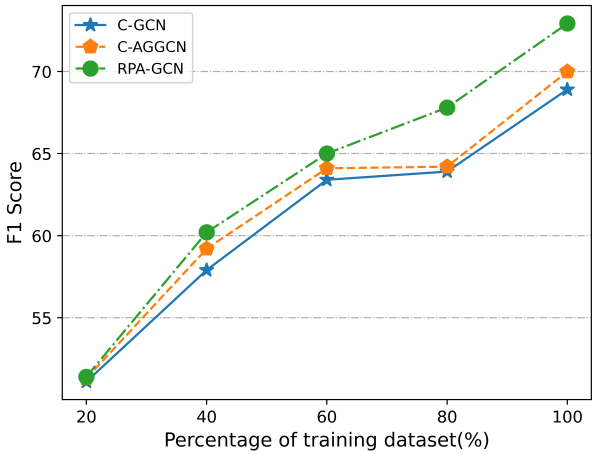

Figure 15. The effect of the number of trainings on F1.

Table 2 presents the results of the ablation tests performed using the RPA-GCN model on the validation set. The relative entity location aware mechanism of the formulation in Section 3.2.1 contributes about 1.8% of the F1 score, while the GAT network contributes about 0.4% and the bidirectional GCN relationship direction-aware mechanism contributes about 1.2%.

Table 2. Ablation analysis.

Model	F1
Final Model	72.68
- Position aware	70.88
- Graph attention networks	70.41
- Bi-directional GCN	69.21
- All of the above	67.58

5. Analysis of Complex Cause-and-Effect Sentences

5.1. Complex Relational Data

To test the effectiveness of our head-to-tail annotation and RPA-GCN model for solving multi-causal and chain-causal problems, we combed the complex causal sentences in the original corpus, as shown in Table 3.

Table 3. Complex causal datasets.

Dataset	Chain of Cause and Effect	Multi-Relationship Causation
All	244	65
Train set	147	39
Validation set	49	13
Test set	48	13

5.2. Model Performance Analysis

Our head-to-tail labeling method converts the relational-extraction task into automatic labeling by parsing all causal pairs through the scoring function S_{final}. For error analysis, we analyzed the confusion matrix of cause label "C" (including "C_h" to "C_t"), effect label "E" (including "E_h" to "E_t"), and chain cause outcome label "CE" (including "CE_h" to "CE_t") on the data of the complex causal corpus, as shown in Figures 16 and 17, and the baseline model for comparison was BI-LSTM+CRF [32].

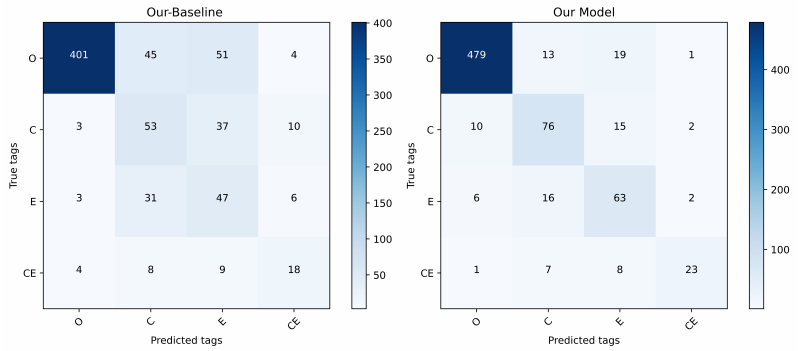

Figure 16. Contrasting models in testing label distributions for linked causality.

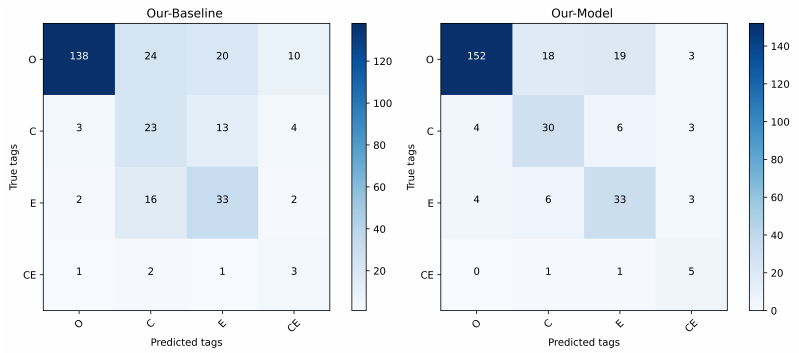

Figure 17. Contrasting models in testing multi-causal label distributions.

As can be seen from Figures 16 and 17, due to the small amount of training data, most of the benchmark model's errors occur in the confusion of entity positions of cause and effect. Our model has the highest automatic labeling accuracy for labels C, E, and CE. With the small amount of training data, the weak character-position-difference representation in text sentences can be compensated to some extent by capturing the relative positions between objects and the bidirectional GCN relationship-direction-aware mechanism. To analyze the impact of different numbers of graph-dependent nodes on the baseline model, further performance analysis was conducted for complex sentences of different lengths.

Figure 18 shows the F1 scores of the three models for different sentence lengths. As the sentence length increases and the dependency graph includes more nodes, the performance of all the models decreases significantly with longer sentences; the information-acquisition performance also decreases. The results show that RPA-GCN outperforms the baseline model under different settings and can obtain better cause–effect correlation information in the dependency graph. Furthermore, our location-encoding strategy reduces the weak correlations from the graph-dependency tree and enhances the associations between causal entities. To clearly represent the improvement of the model performance from the GAT-network-fused entity-location awareness, Figure 19 shows the extraction results of our model and the baseline model [19,32] based on some test sentences.

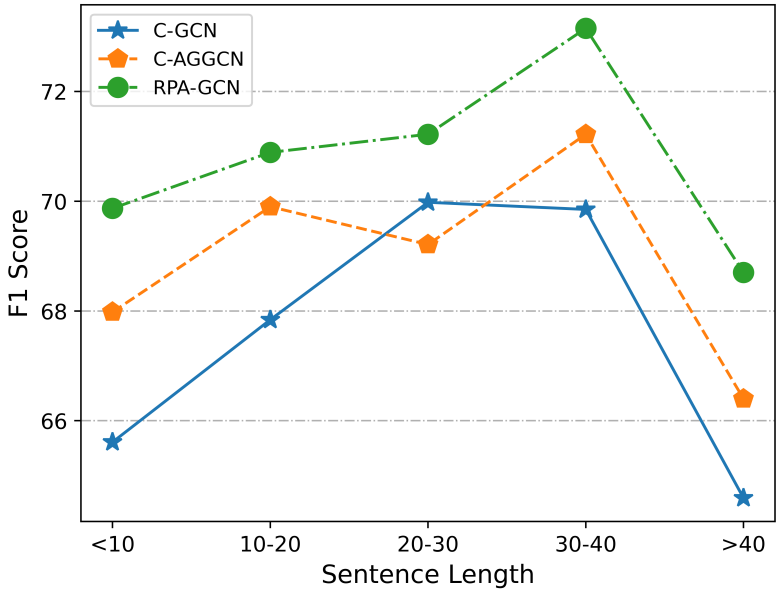

Figure 18. The effects of different sentence lengths on F1.

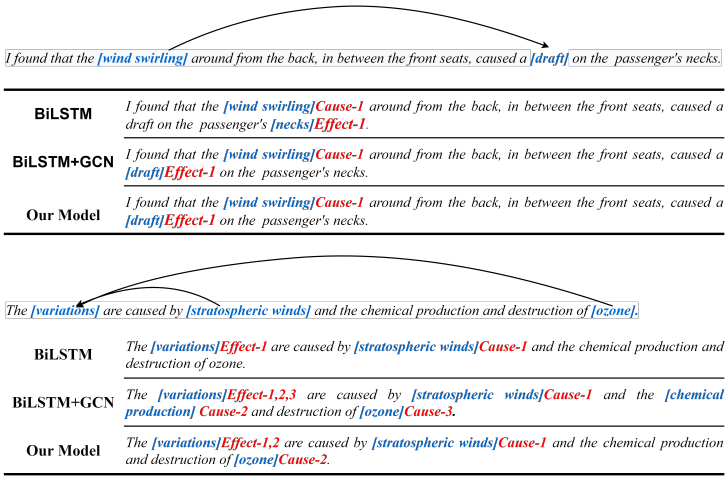

Figure 19. The automatic labeling result of the test sentence pattern. The blue word indicates the causal entity after model labeling, and Cause/Effect-X indicates the X-th causal pair in the sentence pattern.

The first sentence example shows a situation in which the distance between the causal pair entities is long. It is very difficult for the Bi-LSTM [32] to extract relationships, and it cannot build on long syntactic dependencies to analyze the subject and object. In contrast, the GCN model contains contextual information, and it can capture the long-distance dependencies at the syntactic level and directly establish dependency connections between causal pairs, making up for the LSTM's lack of global information extraction ability.

The second sentence example shows a case of chained causality, in which one result corresponds to multiple causes. As the extraction results show, Bi-LSTM+GCN [19] can capture long-range dependencies, but the accuracy of extracting cause and effect on the location of entities is not high enough. For example, the relationship orientation of "chemical production" and "destruction" are both "ozone" rather than a simple juxtaposition cause relationship. Our model benefits from the effective pruning of graph dependencies by entity-location awareness, and this effectively solves the problems of complex cause–effect relationships and long-distance-relationship sentences. We also explored the impact of the

distance of relational entity pairs on the model performance, and the results are shown in Table 4.

Table 4. Effects of different causal entity distances on F1.

	Method		
Distance	Bi-LSTM [32]	Bi-LSTM+GCN [19]	Our Model
<5	0.758	0.868	0.896
[5, 10]	0.391	0.674	0.714
>10	0.219	0.371	0.589

It can be seen that by dividing the distances between causal entities into three classes (<5, [5,10], >10), our model clearly outperforms the baseline in both long- and short-distance relationships. The performance advantage of our model is more obvious when the distance between causal-entity pairs is larger. This is due to the enhancement of the relative-entity-location perception on the causal semantic features.

6. Conclusions

In this paper, we introduce a graph network model incorporating entity location-aware and attention mechanisms for causal relationship extraction and a novel causal relationship labeling scheme that contains head and tail nodes of causal entities and relational words used to jointly extract entity relationships, which is more advantageous in solving complex causal relationship problems without complex feature engineering. Experiments are conducted on sentence-level relation extraction, and the results show that our method obtains the optimal F1 values. Of course, there are some drawbacks and shortcomings in this paper, as there are few datasets for public evaluation of causal relation extraction. We use the extended SemEval 2010 task 8 dataset and Altlex dataset, but the amount of experimental data is still not significant, so the next exploration is to alleviate the data sparsity problem by data augmentation and weak supervision.

Author Contributions: Y.C. conceived and designed the experiment, Y.C. conducted the experiments; Y.W. contributed materials; Y.C. and W.W. conducted the formal analysis. Y.C., J.H. wrote, reviewed, and edited the paper, Y.C., W.W., B.H. revised the manuscript and supervised. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results received funding from the National Key *R&D* Program of China under Grant Agreement Grant No. 2020AAA0109300.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data underlying this article are available in the article.

Acknowledgments: The authors would like to thank all of the anonymous reviewers and editors for their helpful suggestions for the improvement of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RGP-GCN	relation position and attention-graph convolutional networks
GAT	graph attention network
Bi-GCN	bi-directional graph convolutional network
CNN	convolutional neural network
Bi-LSTM	bi-directional long short-term memory

References

1. Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* **2019**, *62*, 54–60. [CrossRef]
2. Zybin, S.; Bielozorova, Y. Risk-based decision-making system for information processing systems. *Int. J. Inf. Technol. Comput. Sci.* **2021**, 13, 1–18. [CrossRef]
3. Young, I.J.B.; Luz, S.; Lone, N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *Int. J. Med. Inform.* **2019**, *132*, 103971. [CrossRef] [PubMed]
4. Jones, N.D.; Azzam, T.; Wanzer, D.L.; Skousen, D.; Knight, C.; Sabarre, N. Enhancing the effectiveness of logic models. *Am. J. Eval.* **2020**, *41*, 452–470. [CrossRef]
5. Jun, E.J.; Bautista, A.R.; Nunez, M.D.; Allen, D.C.; Tak, J.H.; Alvarez, E.; Basso, M.A. Causal role for the primate superior colliculus in the computation of evidence for per-ceptual decisions. *Nat. Neurosci.* **2021**, *24*, 1121–1131. [CrossRef] [PubMed]
6. Dasgupta, T.; Saha, R.; Dey, L.; Naskar, A. 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, Melbourne, Australia, 12–14 July 2018.
7. Fu, J.; Liu, Z.; Liu, W.; Guo, Q. Using dual-layer CRFs for event causal relation extraction. *IEICE Electron. Express* **2011**, *8*, 306–310. [CrossRef]
8. Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; Chang, Y. A novel cascade binary tagging framework for relational triple extraction. *arXiv* **2019**, arXiv:1909.03227.
9. Garcia D. COATIS, an NLP system to locate expressions of actions connected by causality links. In *International Conference on Knowledge Engineering and Knowledge Management*; Springer: Berlin/Heidelberg, Germany, 1997; pp. 347–352.
10. Radinsky, K.; Davidovich, S.; Markovitch, S. Learning causality for news events prediction. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; Association for Computing Machinery: New York, NY, USA, 2012; pp. 909–918. [CrossRef]
11. Zhao, S.; Liu, T.; Zhao, S.; Chen, Y.; Nie, J.-Y. Event causality extraction based on connectives analysis. *Neurocomputing* **2016**, *173*, 1943–1950. [CrossRef]
12. Kim, H.D.; Castellanos, M.; Hsu, M. Mining causal topics in text data: Iterative topic modeling with time series feedback. In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, San Francisco, CA, USA, 27 October–1 November 2013.
13. Lin, Z.; Kan, M.-Y.; Ng, H.T. Recognizing implicit discourse relations in the Penn Discourse Treebank. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009.
14. Wang, L.; Cao, Z.; de Melo, G.; Liu, Z. Relation classification via multi-level attention cnns. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016.
15. Li, P.; Mao, K. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Syst. Appl.* **2019**, *115*, 512–523. [CrossRef]
16. Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; Jin, Z. Classifying relations via long short term memory networks along shortest dependency paths. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015.
17. Zhao, S.; Wang, Q.; Massung, S.; Qin, B.; Liu, T.; Wang, B.; Zhai, C. Constructing and embedding abstract event causality networks from text snippets. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, UK, 6–10 February 2017; pp. 335–344.
18. Li, Z.; Li, Q.; Zou, X.; Ren, J. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing* **2021**, *423*, 207–219. [CrossRef]
19. Zhang, Y.; Qi, P.; Manning, C.D. Graph convolution over pruned dependency trees improves relation extraction. *arXiv* **2018**, arXiv:180910185.
20. Xu, J.; Zuo, W.; Liang, S.; Wang, Y. Causal relation extraction based on graph attention network. *Comput. Res. Dev.* **2020**, *57*, 159. (In Chinese)
21. Dai, D.; Xiao, X.; Lyu, Y.; Dou, S.; She, Q.; Wang, H. Joint extraction of entities and overlapping relations using position-attentive sequence labeling. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
22. Dixit, K.; Al-Onaizan, Y. Span-level model for relation extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
23. Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; Manning, C.D. Position-aware attention and supervised data improve slot filling. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017.
24. de Marneffe, M.-C.; Manning, C.D. *Stanford Typed Dependencies Manual*; Technical report; Stanford University: Stanford, CA, USA, 2008.
25. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:160902907.
26. Saha, S.; Kumar, K.A. Emoji Prediction Using Emerging Machine Learning Classifiers for Text-based Communication. *J. Math. Sci. Comput.* **2022**, *1*, 37–43. [CrossRef]

27. Hendrickx, I.; Kim, S.N.; Kozareva, Z.; Nakov, P.; Séaghdha, D.Ó.; Padó, S.; Pennacchiotti, M.; Romano, L.; Szpakowicz, S. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv* **2019**, arXiv:191110422.
28. Hidey, C.; McKeown, K. Identifying causal relations using parallel Wikipedia articles. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1424–1433.
29. Zheng, Y.; Zuo, X.; Zuo, W.; Liang, S.; Wang, Y. Bi-LSTM+GCN Causal Relationship Extraction Based on Time Relationship. *J. Jilin Univ. (Sci. Ed.)* **2021**, *59*, 643–648. (In Chinese)
30. Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; McClosky, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 23–24 June 2014.
31. Guo, Z.; Zhang, Y.; Lu, W. Attention guided graph convolutional networks for relation extraction. *arXiv* **2019**, arXiv:190607510.
32. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:150801991.