Local-to-Global Causal Reasoning for Cross-Document Relation Extraction

Haoran Wu, Xiuyi Chen, Zefa Hu, Jing Shi, Shuang Xu, and Bo Xu

***Abstract*—Cross-document relation extraction (RE), as an extension of information extraction, requires integrating information from multiple documents retrieved from open domains with a large number of irrelevant or confusing noisy texts. Previous studies focus on the attention mechanism to construct the connection between different text features through semantic similarity. However, similarity-based methods cannot distinguish valid information from highly similar retrieved documents well. How to design an effective algorithm to implement aggregated reasoning in confusing information with similar features still remains an open issue. To address this problem, we design a novel local-to-global causal reasoning (LGCR) network for cross-document RE, which enables efficient distinguishing, filtering and global reasoning on complex information from a causal perspective. Specifically, we propose a local causal estimation algorithm to estimate the causal effect, which is the first trial to use the causal reasoning independent of feature similarity to distinguish between confusing and valid information in cross-document RE. Furthermore, based on the causal effect, we propose a causality guided global reasoning algorithm to filter the confusing information and achieve global reasoning. Experimental results under the closed and the open settings of the large-scale dataset CodRED demonstrate our LGCR network significantly outperforms the state-of-the-art methods and validate the effectiveness of causal reasoning in confusing information processing.**

***Index Terms*—Causal reasoning, cross document, graph reasoning, relation extraction (RE).**

I. Introduction

RELATION extraction (RE) aims to identify semantic relations between entities from unstructured text and plays a crucial role in knowledge acquisition and application systems [1]–[3]. Existing works mainly study this task from two aspects: sentence-level RE [4]–[6] and document-level RE [7]–[9]. Sentence-level RE focuses on entity relations within a sentence, while document-level RE requires aggregation across sentences in a document. Recently, researchers find that there are over 57.6% of relational facts in Wikipedia whose head and tail entities are distributed in different documents [10], [11]. Therefore, extracting relational information in the cross-document scenario has broad practical requirements and is the key to the further development of the knowledge acquisition system.

As shown in Fig. 1, cross-document RE retrieves relevant documents from Wikipedia based on the given entity pair $\{e_h, e_t\}$, and builds multiple reasoning paths for global reasoning. And then the cross-document RE system determines the true relation between two entities by jointly reasoning among multiple reasoning paths. Intuitively, graph structure is a good way to aggregate global information from reasoning paths. And on the document-level RE, graph-based methods achieve SOTA performance [12]–[15]. However, since the documents in cross-document RE are retrieved from Wikipedia, entities are not relevant in some paths. In Fig. 1, we can judge that the relation between two entities is "Child" from the first path, but there is no relation between the two entities in the second path. The path that can be used to determine the relation between entities is the valid path, while the path that has no association between entities is the confusing path. In this case, the neighborhood aggregation propagation of graph neural networks will cause confusion between valid and confusing paths. Unfortunately, the retrieved documents are superficially similar, and there is no clear boundary between the valid paths and confusing paths. Therefore, traditional similarity-based methods cannot distinguish confusing information, and a feature similarity independent method is needed to judge the validity of semantic units in the reasoning paths.

Recently, causal reasoning combined with deep learning has shown strong advantages in many fields [16], [17]. One goal of causal reasoning is to discover causal associations between input features and outcomes. Compared with the similarity-based association, causal association reduces the perplexity of the apparent correlation of features and is therefore more stable and robust [18]. It suggests that we can measure the effectiveness of the information contained in each semantic unit by its causal association with outcomes. To estimate the causal effect, a common method is to exploit the assumption of conditional ignorability [19], which obtains *individual treatment effect* (ITE) by comparing treatments and controls to estimate causal effects independent of the remaining variables. Inspired by this, in the face of confusing reasoning paths with irrelevant information, we can judge the validity of the information from the perspective of causal association with the outcomes

Manuscript received October 20, 2022; revised November 18, 2022 and February 10, 2023; accepted March 11, 2023. This work was supported in part by the National Key Research and Development Program of China (2022ZD0116405), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA27030300), and the Key Research Program of the Chinese Academy of Sciences (ZDBS-SSW-JSC006). Recommended by Associate Editor Nian Zhang. *(Corresponding author: Haoran Wu.)*

Citation: H. R. Wu, X. Y. Chen, Z. F. Hu, J. Shi, S. Xu, and B. Xu, "Local-to-global causal reasoning for cross-document relation extraction," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 7, pp. 1608–1621, Jul. 2023.

The authors are with Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. H. R. Wu, X. Y. Chen, Z. F. Hu, and B. Xu are also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: wuhaoran2018@ia.ac.cn; chenxiuyi2017@ia.ac.cn; huzefa2018@ia.ac.cn; shijing2014@ia.ac.cn; shuang.xu@ia.ac.cn; xubo@ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.123540

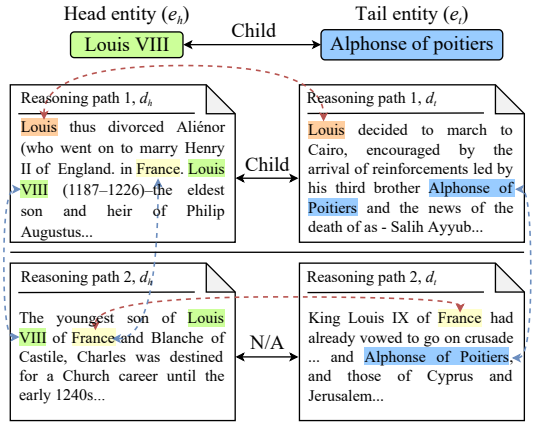

Fig. 1. An example of cross-document RE. A reasoning path consists of two documents $\{d_h, d_t\}$ with head and tail entities, respectively. In a reasoning path, two documents are connected by co-occurring entities, which are bridging entities connected by red dotted lines (such as "Louis" in the reasoning path 1). Bridging entities also exist between reasoning paths, which are connected by blue dotted lines and construct the global connection. Reasoning path 1 is a valid path, "N/A" in reasoning path 2 indicates that there is no relation between the head and tail entities, and it is a confusing path. Both reasoning paths and entity mentions in the figure are semantic units used for reasoning.

rather than the similarity (such as attention scores) between features.

In this paper, we propose a novel method termed as local-to-global causal reasoning (LGCR) network to solve the problem of aggregated reasoning in confusing information for cross-document RE. Specifically, we first use a local relation reasoning module to identify the relation between the head and tail entities in each reasoning path. Based on this, we propose a local causal estimation algorithm. It implements counterfactual treatment by masking the expression of specific semantic unit while keeping other features unchanged. The causal effect of the semantic unit is then estimated by comparing the predicted distributions under the treatment and control conditions, where the control indicates that all features are kept constant. It can clearly indicate how different semantic units contribute to the establishment of the relation in the reasoning path based on causality rather than correlation. Then, we propose a causality guided global reasoning algorithm to solve the problem of global reasoning under the interference of confusing information. We construct a global reasoning graph for cross-document RE based on the co-occurrence of entity mentions and the structure of reasoning paths. With the graph, the algorithm uses relative causal association calculated by local causal effect to control the message propagation ability between nodes. To minimize the influence of confusing semantic unit nodes in the graph, we truncate edges with relative association less than a threshold and use a reparameterization algorithm to solve the gradient propagation problem. Experimental results on the closed and the open settings on CodRED demonstrate that our method substantially outperforms previous models.

The main contributions of this work can be summarized as follows:

1) We propose a novel LGCR network to aggregate reasoning in confusing information with similar features from a causal perspective. To our knowledge, we are the first to use the causal reasoning to solve the confusing information filtering problem in cross-document RE.

2) We propose a local causal estimation algorithm to estimate the causal effect for each semantic unit to distinguish between confusing and valid information. Based on the causal effect, we propose a causality guided global reasoning algorithm to filter confusing information and realize aggregation global reasoning.

3) Experiments on two settings on a large-scale dataset CodRED show that our method significantly outperforms existing state-of-the-art methods. Analytical experiments show that our method successfully discriminates target-related and irrelevant semantic units and filter confusing information in global reasoning, which demonstrates the effectiveness of causal reasoning in the processing of confusing information.

II. Related Work

1) Relation Extraction: The earliest RE is mainly performed within a sentence [20]–[23]. Subsequently, to expand the application scope of RE and improve accuracy, document-level RE has been fully studied in recent years [24]–[26]. In the document-level RE, sequential [27] or graph structures [15] are usually used to model documents and adaptive thresholds [9] are applied to determine whether relations exist between entities. In the graph-based method, Zeng *et al*. [14] build a mention graph and an entity graph for continuous reasoning, and Xu *et al*. [15] optimize the structure of the graph through iterative reconstruction. To extract accurate knowledge from a wider range of data, Yao *et al*. [11] start the exploration of cross-document RE. However, most of the existing methods cannot handle the confusing information filtering problem in complex cross-document scenarios. To this end, we focus on the modeling of cross-document RE with causal methods in this paper.

2) Information Filtering: Filtering the input information can effectively improve the performance of the model [28]. As we need to filter out confusing paths from retrieved reasoning paths, how to filter out valuable information from the massive information has received extensive attention [29]–[31]. The challenge is to deal with confusing and incomplete data [32]. Regularization constraint is an effective method to deal with incomplete data. It can be used to describe the temporal patterns in a high-dimensional and sparse tensor to assist in the prediction of incomplete data [33]; or to enhance the robustness and stability of the algorithm [34]. For confusing data, Hu *et al*. [35] propose an inductive clustering algorithm based on co-occur feature to identify confusing clusters in attributed graph, and Shrikumar *et al*. [36] discover the valid information in the confusing data based on propagating activation differences. Differently from the previous work, we use causal reasoning to distinguish valid information from confusing information.

3) Causal Reasoning Methods: Typical causal reasoning methods aim to discover and eliminate confusing factors and

spurious associations by learning balanced weights of samples [37]–[39]. As a branch of causal inference, counterfactual reasoning is used to identify causal associations between treatments and outcomes [19], and is mainly used in recommendation systems [40] and medical treatment [41]. In NLP, it is often used to implement data augmentation to enhance the generalization of the model [42], [43]. For wider application, Zhu *et al*. [44] use counterfactual contrast in dialogue generation to improve the quality of responses, and Wu *et al*. [16] explore extracting key information from documents based on causal associations. In this paper, we are the first to use the causal association between different semantic units and outcomes to guide the global reasoning of information in cross-document RE.

III. Methodology

A. Problem Formulation

Firstly, for convenience, the descriptions of some important symbols used in this paper are summarized in Table I. Following the previous work [11], in cross-document RE, we mainly focus on the reasoning between multiple complementary reasoning paths. A reasoning path can be a document pair $\{d_h, d_t\}$ containing the head entity and tail entity respectively, connected by bridging entities. So we formulate the cross-document RE as follows. Given two entities $\{e_h, e_t\}$ and a set of reasoning paths $\mathcal{D} = \{d_h^i, d_t^i\}_{i=1}^N$ retrieved from Wikipedia, the goal of this paper is to reason within the complex information of these paths and determine the global relation $R_{\mathcal{D}}$ between head and tail entities.

TABLE I
The Description of Some Important Symbols

Symbol	Description
e_h, e_t	The head and tail entities for RE.
d_h, d_t	The documents containing head and tail entities.
$\mathcal{D}$	A set of reasoning paths.
$R_{\mathcal{D}}$	The global relation between head and tail entities.
X, Y, T	The causal, result and treatment in LEC.
u	The semantic unit that requires causal estimation.
x_u	The feature of semantic unit u.
$\boldsymbol{h}_p^i, \boldsymbol{h}_h^i, \boldsymbol{h}_t^i$	The encoder representation of the reasoning path, head entity and tail entity of reasoning path i.
$\bar{\mathfrak{h}}_p^i, \tilde{\mathfrak{h}}_p^i$	The local and global representation of reasoning path i.
$\hat{\mathfrak{h}}_u^i$	The treatment representation of semantic unit u in reasoning path i.
$\boldsymbol{I_u}$	The ITE of semantic unit u.
$\mathfrak{g}_a^m$	The representation of node a in m-th layer in CGGC.
$\tilde{\mathfrak{g}}_a$	The global representation of node a obtained from CGGC.
$r_{a,b}$	The relative causal effect between the node a and b in CGGC.
$s_{a,b}^*$	The connection selector of the edge between the node a and b in CGGC.
$\bar{\boldsymbol{P}}_i, \tilde{\boldsymbol{P}}_i, \hat{\boldsymbol{P}}_i^u$	The relation distribution of reasoning path i calculated by local representation $\bar{\mathfrak{h}}_p^i$, global representation $\tilde{\mathfrak{h}}_p^i$ and treatment representation $\hat{\mathfrak{h}}_u^i$.
$\boldsymbol{P}^{\mathcal{D}}$	The final distribution of reasoning path set $\mathcal{D}$.

In this paper, we use causal estimation to evaluate the local causal effect of each semantic unit. This paper has four kinds of semantic units: the reasoning path unit corresponding to each path, the mention unit corresponding to the head entity, the tail entity, and the bridging entity. Based on the information of these four kinds of semantic units, we can infer the relation between entities in the cross-document scenario. We define the causality in our method as follows. During the local reasoning, the semantic information of a particular semantic unit contained in the reasoning path is the cause (X), and the prediction of this reasoning path is the result (Y). Therefore, the causality is the mapping from X to Y, and the local causal effect of this semantic unit can be estimated by the contribution of its semantic information to the reasoning result.

B. Local Causal Estimation

1) Leveraging Causal Reasoning: The counterfactual method is used in causal reasoning to answer questions about "what if" to discover causal relationships between treatments and outcomes based on the conditional ignorability [19]. Conditional ignorability means that, given a background variable X, the treatment T and the potential outcomes Y that could occur under the treatment and control (without treatment) are independent of each other. In other words, it requires that the background variable X is the same for the treatment and control when causal estimated. This avoids the different outcome tendencies for the treatments due to the different distributions of X. Therefore, the causal effect of the treatment can be estimated by comparing two outcomes with the same background variable X but with different treatment statuses. Based on this, using counterfactual reasoning requires building a basic model, which can be a nonlinear mapping $Y = f(X)$ from context X to outcome Y. As shown in Fig. 2, in order to estimate the causal association of semantic units u in context X to outcome Y, we need to apply a treatment T_u on X. A treatment $T_u = \{0, 1\}$ in this paper represents the presence or absence of the feature x_u of a specific semantic unit u with the remaining variables X_{T_u} unchanged. This causal effect is usually measured by *individual treatment effect* (ITE) [45], [46]. Specifically, denoting $Y = f(x_u, X_{T_u} | T_u = 0)$ as the basic model for counterfactual reasoning without treatment, the treatment outcome can be denoted as $f(0, X_{T_u} | T_u = 1)$, which indicates that the feature x_u cannot express during the inference. Then, the ITE of feature x_u can be formulated as

$$\mathrm{ITE}(x_u) = f(x_u, X_{T_u} | T_u = 0) - f(0, X_{T_u} | T_u = 1). \quad (1)$$

2) Local Reasoning Module: To estimate the ITE of each semantic unit, we construct a local reasoning module as the

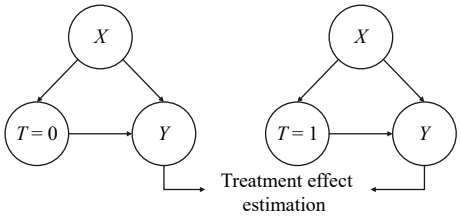

Fig. 2. Estimate the causal effect of T on Y by treatment-control experiment.

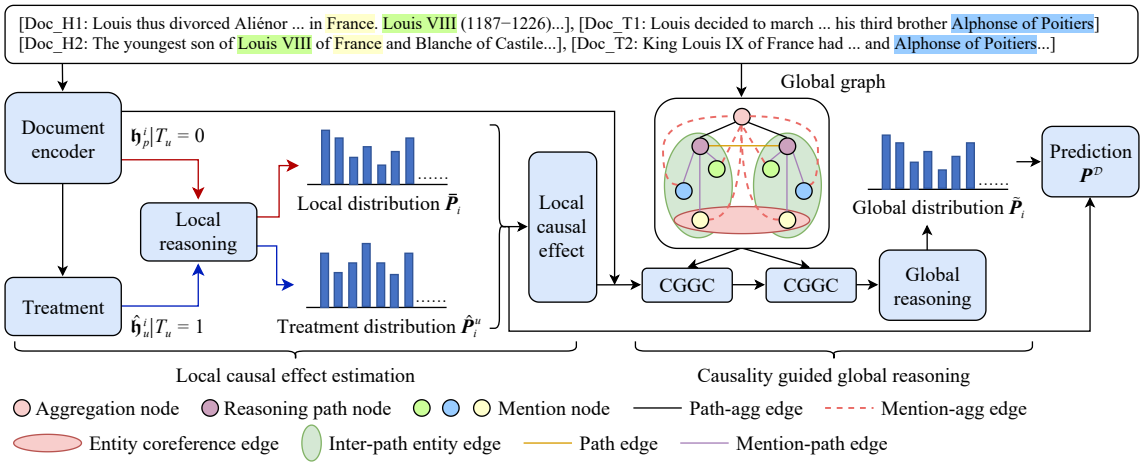

Fig. 3. The overview of the LGCR. The red arrows represent local reasoning without the treatment, and the blue arrows represent the local reasoning after the treatment. We calculate local effects by comparing the difference between the local distribution and treatment distribution. In the global graph, we omit some types of edges for a clear presentation. The three types of mention nodes represent head, tail and bridging entities respectively. The edges represented by circles represent the fully connected relationship between the nodes.

basic model for counterfactual reasoning, which aims to identify the relation of entity pair $\{e_h, e_t\}$ within a reasoning path i as $X_i = \{d_h^i, d_t^i\}$ based on the text fragments and entity mentions. We adopt BERT [47] as the document encoder in Fig. 3. BERT is a language model based on the deep transformer [48] structure and pre-trained on a large-scale corpus by the methods of masked language model and next sentence prediction. Specifically, given a sequence of tokens as input X, BERT first converts it to the token embedding, the segment embedding, the position embedding respectively and sums them up. Then, BERT encodes the input embedding with M-stacked transformer blocks. For the m-th transformer block, the representation $\boldsymbol{H}^m$ is updated by multi-head self-attention as

$$\boldsymbol{Q} = \boldsymbol{H}^{m-1}\boldsymbol{W}_Q^m; \boldsymbol{K} = \boldsymbol{H}^{m-1}\boldsymbol{W}_K^m; \boldsymbol{V} = \boldsymbol{H}^{m-1}\boldsymbol{W}_V^m \tag{2}$$

$$\boldsymbol{H}^m = \text{FeedForward}\left(\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}'}{\sqrt{\dim_K}}\right)\boldsymbol{V}\right) \tag{3}$$

where $\boldsymbol{W}_Q^m$, $\boldsymbol{W}_K^m$, $\boldsymbol{W}_V^m$ are learnable matrix, $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ represent the query, key and value in scaled dot-product attention respectively, and $\dim_K$ represents the dimension of $\boldsymbol{K}$.

To use BERT, for each reasoning path, documents are first concatenated and tokenized. Then, we insert special tokens before and after all relevant entity mentions ($\{[\text{UNUSED}k]\}_{k=1,2}$ for head entity, $\{[\text{UNUSED}k]\}_{k=3,4}$ for tail entity and others for remaining entity mentions). Since the length of documents in each reasoning path far exceeds the limit of BERT, we only extract fragments surrounding the head and tail entities for experiments.

After processing the context input X_i of i-th reasoning path, we leverage the BERT encoder to obtain the representation of each token as

$$\boldsymbol{H}_i = \text{BERT}(X_i) = \{\boldsymbol{h}_{cls}^i, \boldsymbol{h}_0^i, \ldots, \boldsymbol{h}_n^i\} \tag{4}$$

where n is the number of tokens in i-th reasoning path. We use the "[CLS]" token representation of the last layer in BERT as the overall representation of path i, denoted as $\boldsymbol{h}_p^i$. For the head and tail entities in reasoning path i, we use the average representation of all the tokens they contain, denoted as $\boldsymbol{h}_h^i$ and $\boldsymbol{h}_t^i$. Then, we concatenate the three as the final local representation of path i, denoted as

$$\bar{\mathfrak{h}}_p^i = [\boldsymbol{h}_p^i; \boldsymbol{h}_h^i; \boldsymbol{h}_t^i]. \tag{5}$$

Finally, we use a fully connected layer as the local reasoning module in Fig. 3 to calculate the local distribution $\bar{\boldsymbol{P}}_i$ of entity pair in reasoning path i as

$$\bar{\boldsymbol{P}}_i = \bar{\mathfrak{h}}_p^i \boldsymbol{W}_l + \boldsymbol{b}_l \tag{6}$$

where $\boldsymbol{W}_l$ is the learnable matrix and $\boldsymbol{b}_l$ is the bias.

3) Local Causal Reasoning: The process of local causal reasoning is shown in the left half of Fig. 3. Taking the local reasoning module of (6) as the basic causal reasoning model $Y = f(X)$, the causal effect of each semantic unit u on the outcome Y can be estimated with the treatment. In this paper, the treatment means whether the representation of feature x_u of the specific semantic unit u is included in the local representation $\bar{\mathfrak{h}}_p^i$ of path i during the local reasoning. The treatment representation of semantic unit u in reasoning path i is denoted as $\hat{\mathfrak{h}}_u^i$. For the following graph reasoning, we need to estimate the ITE for the four types of semantic units, respectively. For the treatment representation $\hat{\mathfrak{h}}_p^i$ of reasoning path i, $\hat{\mathfrak{h}}_h^i$ of head entity in reasoning path i and $\hat{\mathfrak{h}}_t^i$ of tail entity in reasoning path i, they can be simply implemented by replacing the corresponding representation in $\bar{\mathfrak{h}}_p^i$ with a zero vector in (5) as

$$\hat{\mathfrak{h}}_p^i = [0; \boldsymbol{h}_h^i; \boldsymbol{h}_t^i]; \hat{\mathfrak{h}}_h^i = [\boldsymbol{h}_p^i; 0; \boldsymbol{h}_t^i]; \hat{\mathfrak{h}}_t^i = [\boldsymbol{h}_p^i; \boldsymbol{h}_h^i; 0]. \tag{7}$$

And for other entity mention j in reasoning path i, the same as the representation of head and tail entities, we use the average representation of all the tokens it contains, denoted as $\boldsymbol{h}_o^{i,j}$. To eliminate the influence of entity mention j, we subtract it from the reasoning path representation $\boldsymbol{h}_p^i$ of path i in $\bar{\mathfrak{h}}_p^i$ to obtain the treatment representation as

$$\hat{\mathfrak{h}}_o^{i,j} = [\boldsymbol{h}_p^i - \boldsymbol{h}_o^{i,j}; \boldsymbol{h}_h^i; \boldsymbol{h}_t^i]. \tag{8}$$

Then, we can calculate the treatment distribution $\hat{\boldsymbol{P}}_i^u$ of semantic unit u based on the treatment representation $\hat{\mathfrak{h}}_u^i$ with the same fully connected layer as the local reasoning module in (6):

$$\hat{\boldsymbol{P}}_i^u = \hat{\mathfrak{h}}_u^i \boldsymbol{W}_l + \boldsymbol{b}_l \tag{9}$$

where $\hat{\mathfrak{h}}_u^i \in \{\hat{\mathfrak{h}}_p^i, \hat{\mathfrak{h}}_h^i, \hat{\mathfrak{h}}_t^i, \hat{\mathfrak{h}}_o^{i,j}\}$. And the ITE of each semantic unit u is derived by comparing the local distribution with the treatment distribution in the way of (1) as

$$\boldsymbol{I}_u = \sigma\left(\text{Softmax}(\bar{\boldsymbol{P}}_i) - \text{Softmax}(\hat{\boldsymbol{P}}_i^u)\right) \tag{10}$$

where σ is the RELU activation function to ensure that only the ITEs of categories with higher scores in local reasoning are captured.

C. Causality Guided Global Reasoning

1) Global Graph Construction: To aggregate local information over multiple reasoning paths, we first construct a global heterogeneous graph. As shown in the right half of Fig. 3, the graph has three types of nodes:

- **Mention Node**, which represents the entity mentions.
- **Reasoning Path Node**, which is used to represent the overall information of each reasoning path.
- **Aggregation Node**[1], which is a virtual extra node that aggregates reasoning information for all paths and alleviates the long-distance dependency problem on the entire graph.

Then, according to natural understanding, we define the edges for six relations.

- **Entity Coreference Edge:** Mentions referring to the same entity are fully connected.
- **Inter-Path Entity Edge:** Mentions contained in the same reasoning path are fully connected.
- **Path Edge:** All the reasoning path nodes are fully connected.
- **Mention-Path Edge:** Mention nodes are connected to their corresponding path nodes.
- **Mention-Agg Edge:** All the mention nodes are connected to the aggregation node.
- **Path-Agg Edge:** All the path nodes are connected to the aggregation node.

With the connections above, we build the hierarchical relations within all reasoning paths and an information exchange bridge for entities between different reasoning paths, so that the reasoning process is no longer limited to the single reasoning path. The connections to the aggregation node allow all semantic units to interact through two-hop neighbors, enabling effective information integration. And the connections between entities and reasoning paths model the hierarchical semantic structure within the document.

2) Causality Guided Graph Convolution: After constructing the global graph, we design a causality guided graph convolution (CGGC) layer to aggregate the information of multi reasoning paths. We first define the input feature $\mathfrak{g}_a^0$ for node a in CGGC. Specifically, we use the representation of the corresponding entity mention in $\{\boldsymbol{h}_h^i, \boldsymbol{h}_t^i, \boldsymbol{h}_o^{i,j}\}$ for the mention node, use the representation of the reasoning path $\boldsymbol{h}_p^i$ for the reasoning path node, and use the average representation of all semantic unit for the aggregation node.

On this basis, the local causal effects obtained in Section III-B-3) are introduced to control the ability of message propagation between nodes. The relative causal effect between two nodes a and b can be calculated with normalization as

$$r_{a,b} = \frac{\exp(\boldsymbol{I}_a \cdot \boldsymbol{I}_b')}{\sum_{v \in \mathcal{N}_e(a)} \exp(\boldsymbol{I}_a \cdot \boldsymbol{I}_v')} \tag{11}$$

where $\boldsymbol{I}_b'$ represents the transpose of $\boldsymbol{I}_b$ and $\mathcal{N}_e(a)$ represents all one-hop neighbors of node a in e-th type edge.

A low ITE score means that the information contained in the node has little influence on the final outcomes. Therefore, to eliminate the influence of these noises, we truncate the interactive connection between these irrelevant nodes and their neighbors. Specifically, we perform an adaptive sampling on the graph based on the relative causal effect $r_{a,b}$. When $r_{a,b}$ is smaller than a threshold, we truncate the connection between nodes a and b. However, this operation is not differentiable during the back-propagation process to optimize the model. Thus, we use the reparameterization algorithm [49], [50] to bypass the problem. During training, we first use $r_{a,b}$ to construct a Bernoulli distribution as

$$\{\pi_1 := r_{a,b}, \pi_0 := 1 - r_{a,b}\} \tag{12}$$

where π_1 and π_0 represent the probabilities that the connection between nodes a and b is preserved or truncated respectively. Then, we adopt Gumbel-Softmax approach to generate the differentiable selective probability $s_{a,b}$ as

$$s_{a,b} = \frac{\exp((\ln(\pi_1) + g_1)/\lambda)}{\sum_{k \in \{0,1\}} \exp((\ln(\pi_k) + g_k)/\lambda)} \tag{13}$$

where λ denotes relaxation temperature, g_0 and g_1 are independent noises sampled from the Gumbel distribution [51]. Then, we can build a connection selector as

$$s_{a,b}^* = \text{detach}(s_{a,b}' - s_{a,b}) + s_{a,b} \tag{14}$$

where "detach" represents that the gradient is truncated at optimization time. $s_{a,b}' = 0$ when $s_{a,b}$ is smaller than the threshold, and otherwise $s_{a,b}' = 1$. In this way, we make $s_{a,b}^* \in \{0, 1\}$ differentiable. Finally, relative causal effects are introduced into the convolution computation for the feature of node a in CGGC as

$$\mathfrak{g}_a^m = \sigma\left(\sum_{e \in \mathcal{E}} \sum_{b \in \mathcal{N}_e(a)} \frac{\mathfrak{g}_b^{m-1} \boldsymbol{W}_e^m \odot r_{a,b}^*}{\sqrt{|\mathcal{N}_e(a)||\mathcal{N}_e(b)|}}\right)$$
$$r_{a,b}^* = r_{a,b} * s_{a,b}^* \tag{15}$$

where $\odot$ means Hadamard product, $\mathcal{E}$ represents different types of edges, $\boldsymbol{W}_e^m$ is the learnable matrix for the e-th type edge of layer m, $\mathfrak{g}_a^m$ is the representation of node a of layer m, and $\sqrt{|\mathcal{N}_e(a)||\mathcal{N}_e(b)|}$ is the normalization constant based on the graph structure.

D. Prediction and Training

After obtaining the graph representation, we concatenate the outputs of each layer, denoted as $\tilde{\mathfrak{g}}_a = [\mathfrak{g}_a^0; \ldots; \mathfrak{g}_a^M]$, where $\mathfrak{g}_a^0$ is

[1] We set the ITE of the aggregation node to 1.

the representation of node a before CGGC and M is the number of CGGC layers. For the representation $\tilde{\mathfrak{h}}_p^i$ of reasoning path i in global reasoning, we use the same structure $\tilde{\mathfrak{h}}_p^i = [\tilde{\mathfrak{g}}_p^i; \tilde{\mathfrak{g}}_h^i; \tilde{\mathfrak{g}}_t^i]$ as the local reasoning module. To obtain the global relation across documents, we first calculate the distribution of local relation for path i based on the global representation by

$$\tilde{\boldsymbol{P}}_i = \tilde{\mathfrak{h}}_p^i \boldsymbol{W}_g + \boldsymbol{b}_g \tag{16}$$

where $\boldsymbol{W}_g$ is the learnable matrix and $\boldsymbol{b}_g$ is the bias. Then, following Alvarez Melis and Jaakkola [52], we use a regularization constraint to enhance the reasoning ability of local causal estimation. Specifically, we calculate the treatment distribution $\hat{\boldsymbol{P}}_i^u$ of each semantic unit u in reasoning path i based on the corresponding treatment representation $\hat{\mathfrak{h}}_u^i$ using the same local reasoning module as (6). Then we calculate the average distribution $\hat{\boldsymbol{P}}_i$ for each reasoning path i as $\hat{\boldsymbol{P}}_i = \frac{1}{V_i}\sum_{u=0}^{V_i} \hat{\boldsymbol{P}}_i^u$, where V_i is the number of semantic units in path i. And as shown in the rightmost part of Fig. 3, we take the sum of the distribution obtained from the three modules as the final distribution of reasoning path i for prediction as

$$\boldsymbol{P}_i = \alpha \bar{\boldsymbol{P}}_i + \beta \tilde{\boldsymbol{P}}_i + \gamma \hat{\boldsymbol{P}}_i \tag{17}$$

where α, β, γ are hyperparameters. Finally, we select the highest score from the predicted distributions of all reasoning paths as the final distribution for each relation as $\boldsymbol{P}^{\mathcal{D}} = \text{Max}\{\boldsymbol{P}_i\}_{i=1}^N$.

For training, due to the existence of confusing data where head and tail entities are irrelevant, we optimize the model by a variant of circle loss [53] with a threshold, which is used to distinguish between positive relations and negative relations. Formally, the loss can be calculated as

$$\mathcal{L} = \ln\left(e^{\theta} + \sum_{r\in\Omega_{neg}} e^{p_r}\right) + \ln\left(e^{-\theta} + \sum_{r\in\Omega_{pos}} e^{-p_r}\right) \tag{18}$$

where p_r is the score of relation r in $\boldsymbol{P}^{\mathcal{D}}$, θ is the threshold, Ω_{neg} and Ω_{pos} are the sets of negative and positive relations respectively. Finally, we choose the relation with the highest score in $\boldsymbol{P}^{\mathcal{D}}$ as the final relation $R_{\mathcal{D}}$. We summarize the overall flow of LGCR into pseudocode as shown in Algorithm 1.

Algorithm 1 LGCR

Input: The text for N reasoning paths $\{X_i\}_{i=1}^N$.
Output: The cross-document relation $R_{\mathcal{D}}$.
for reasoning path $i \in [[N]]$ **do**
  $\boldsymbol{H}_i = \text{BERT}(X_i)$;
  Use (5) to obtain the path representation $\tilde{\mathfrak{h}}_p^i$;
  for semantic unit u in reasoning path i **do**
    Calculate the treatment representation $\hat{\mathfrak{h}}_u^i$;
    Estimate the causal effect $\boldsymbol{I}_u$ based on (10);
  end for
end for
for one node a and its neighbor $b \in \mathcal{N}_e(a)$ **do**
  Use (11) to obtain the relative causal effect $r_{a,b}$;
  Construct the Bernoulli distribution based on $r_{a,b}$;
  Generate the connection selector $s_{a,b}^*$;
  if $s_{a,b}^* = 0$ **then**
    $r_{a,b}^* = 0$;
  end if
end for
Update all nodes' representations based on (15).
Calculate the final distribution $\boldsymbol{P}^{\mathcal{D}}$ based on the representation $\tilde{\mathfrak{h}}_p^i$, $\tilde{\mathfrak{h}}_p^i$ and $\hat{\mathfrak{h}}_u^i$ from three modules.

IV. Experiments

A. Experimental Settings

We conduct experiments on both the closed setting and the open setting of CodRED [11], which is constructed based on Wikipedia. CodRED contains 276 relations types, 4755 positive relational facts, and 13 686 positive reasoning paths, alone with 25 749 N/A relational facts and 197 126 N/A reasoning paths. Therefore, how to find effective information from a large number of confusing paths is a huge challenge. For the closed setting, we use the limited reasoning paths given in the dataset. For the open-setting, we retrieve up to 16 reasoning paths for each given entity pair from Wikipedia with entity count. The entity count is a retrieval rule that multiplies the occurrence number of e_h in d_h and the occurrence of e_t in d_t, and chooses the top k reasoning paths $\{d_{ih}, d_{it}\}_{i=1}^k$. Following previous works [5], we use aggregate precision-recall curves with the area under curve (AUC) and the maximum F1 on the curve and Precision@k (P@K) to evaluate our model in the main experiments.

We implement our method with PyTorch for all codes, HuggingFace for the BERT-based model and DGL for the graph network. The number of CGGC layers is set to 2. The dropout ratio of our model is set to 0.2. For the connection sampling, the relaxation temperature λ is set to 0.2 and the truncation threshold is set to 0.5. For the prediction, the α, β and γ are set to 1.0, 0.8 and 0.01. For the optimization objective, θ is set to 10. During the training, we set the learning rate to 3e–5. And for each experiment, we train the model for only two epochs. Other parameters are consistent with previous work [11]. We conduct our BERT$_{base}$ experiments on RTX 3090 GPU and for the RoBERTa$_{large}$ experiments we use A100-40G GPU.

We compare our proposed method LGCR with the following works:

BERT-pipeline [11], which first uses a BERT-based document-level RE model [8] to extract the relation graph between entities in each document, and then performs cross-document RE on the graph based on bridging entities.

BERT-attn [11], which simultaneously extracts the relations of entities within and across documents. Specifically, it uses BERT to encode each reasoning path, then uses the representation of "[CLS]" token to represent the reasoning path, and finally uses selective attention [5] to aggregate the global information.

CorefBERT-RE [54]. CorefBERT is pre-trained by capturing the coreferential relations in context and is often used in document-level RE. We use the CorefBERT to encode each reasoning path and use selective attention to aggregate the global information.

TABLE II
MAIN RESULTS OF TWO BENCHMARK SETTINGS. RESULTS WITH ‡ ARE REPORTED IN THEIR ORIGINAL PAPERS. RESULTS WITH * ARE WE REPRODUCED FROM THE SOURCE CODE OF THE CORRESPONDING PAPER IN CROSS-DOCUMENT RE

Model	Closed						Open					
	Dev				Test		Dev				Test	
	AUC	F1	P@500	P@1000	AUC	F1	AUC	F1	P@500	P@1000	AUC	F1
BERT-pipeline$^{\ddagger}_{base}$	17.45	30.54	30.60	26.70	18.94	32.29	14.07	26.45	27.00	19.90	16.26	28.70
BERT-attn$^{\ddagger}_{base}$	47.94	51.26	62.80	51.00	47.46	51.02	40.86	47.23	59.00	46.30	39.05	45.06
CorefBERT-RE$^{*}_{base}$	49.91	53.17	65.27	52.95	50.56	54.06	41.87	47.72	57.88	47.45	43.56	49.39
GAIN-BERT$^{*}_{base}$	51.04	53.71	63.07	53.45	50.56	54.47	43.77	47.91	57.68	47.95	44.62	49.01
PAR-DRE-BERT$^{*}_{base}$	55.92	56.29	69.46	57.74	54.62	55.63	45.98	49.16	64.27	49.55	45.80	50.61
Ecrim-BERT$^{*}_{base}$	59.92	59.99	74.65	60.44	59.46	59.43	41.15	47.26	60.88	47.25	43.98	48.39
LGCR-BERT$_{base}$	**63.17**	**61.67**	**76.65**	**61.84**	**61.08**	**60.75**	**51.48**	**52.96**	**70.06**	**52.19**	**50.15**	**53.45**
RoBERTa-attn$^{*}_{large}$	52.24	54.21	67.07	53.95	50.41	54.97	42.60	48.33	60.68	48.15	42.81	48.87
CorefRoBERTa-RE$^{*}_{large}$	51.10	55.74	67.27	55.24	49.43	53.65	41.42	48.27	60.48	48.35	40.38	48.01
GAIN-RoBERTa$^{*}_{large}$	57.26	58.21	70.06	58.04	53.34	57.93	46.25	51.07	61.88	51.15	43.30	50.96
PAR-DRE-RoBERTa$^{*}_{large}$	59.26	59.80	76.25	59.64	55.01	58.58	48.35	52.24	66.67	52.25	46.17	52.32
Ecrim-RoBERTa$^{*}_{large}$	61.83	60.54	76.05	60.95	60.39	61.71	44.19	49.14	60.87	49.35	41.75	49.84
LGCR-RoBERTa$_{large}$	**64.76**	**63.18**	**77.25**	**63.74**	**63.03**	**63.79**	**52.36**	**55.15**	**71.66**	**55.14**	**49.05**	**55.37**

GAIN [14], which uses BERT to encode the context, combines the mention-level and entity-level graph to reasoning in the document. We apply it to cross-document RE by treating reasoning paths as sentences.

PAR-DRE [55], which is the latest improvement of GAIN in dialogue relation extraction, using the position-aware graph attention mechanism to address the distinction of similar topological structures in graph neural network.

Ecrim [56], which filters the confusing information in the reasoning path based on the entity co-occurrence relationship and the similarity of the sentences. Then it uses BERT to encode the context, and builds an association matrix between entities based on the attention mechanism to guide information aggregation.

In addition, we also compare the graph reasoning ability with GCN [57], GAT [58] and Graph-SAGE [59].

B. Experimental Results

1) Main Results: We conduct experiments based on two pretrained models, BERT$_{base}$ and RoBERTa$_{large}$. The main results are shown in Table II. We can see that:

- Our method has significant improvement over the baselines on all metrics in both the closed and the open settings, which demonstrates the effectiveness of our approach to guiding global reasoning through the local causal effect on this task.
- Compared with attention-based methods, graph-based methods can more accurately judge the relation between entity pairs due to their global reasoning ability. And LGCR achieves better performance by filtering confusing semantic unit information.
- Compared with the method of filtering confusing information based on the semantic similarity (Ecrim), LGCR has significant advantages in both the settings. Especially, in the open setting, Ecrim suffers a large amount of confusing information, resulting in a much weaker performance than LGCR, which demonstrates the advantage of local causal effect over similarity in confusing information filtering.
- All RoBERTa$_{large}$-based methods achieve effective improvements compared with BERT$_{base}$-based methods in the closed setting. Compared with other methods, our method LGCR achieves significant improvements on both BERT$_{base}$ and RoBERTa$_{large}$, demonstrating the consistent improvement of LGCR under different parameter numbers.

2) Ablation Studies: To demonstrate the effectiveness of each module we proposed, we conduct ablation experiments by deleting specific modules respectively. In the experiments, "w/o sampling" represents the removal of connection sampling; "w/o RC" represents the removal of regularization constraint of treatment distribution; "w/o LEC" represents the removal of local causal estimation, and the ITE of all semantic units is set to 1; "w/o CGGC" represents the removal of the whole global graph. The experimental results are shown in Table III.

From the experimental results, we can see that all modules contribute to performance improvement. Among all modules, the removal of LCE has the greatest influence on the results. This is because when the ITE of all semantic units is set to 1, the graph will simply aggregate the information of neighbors, which is seriously disturbed by confusing information. This demonstrates the effectiveness of LCE in filtering confusing information. The removal of CGGC also has a great influence on the results, which indicates that global information reasoning is necessary for cross-document RE.

3) Comparisons Between Different Graph Networks for Global Reasoning: To verify the effectiveness of causal effects in graph reasoning, we compare some representative graph neural networks, including GCN [57], GAT [58] and GraphSAGE [59]. During the experiments, we only replace CGGC with the corresponding graph neural network respec-

TABLE III
ABLATION STUDY BY REMOVING THE MAIN COMPONENTS, WHERE "W/O" INDICATES WITHOUT. "SAMPLING", "RC", "LCE", "CGGC" REFER TO CAUSAL CONNECTION SAMPLING, REGULARIZATION CONSTRAINT, LOCAL CAUSAL ESTIMATION AND CAUSALITY GUIDED GRAPH CONVOLUTION

Model	Closed				Open			
	AUC	F1	P@500	P@1000	AUC	F1	P@500	P@1000
LGCR	**63.17**	**61.67**	**76.65**	**61.84**	**51.48**	**52.96**	**70.06**	**52.19**
w/o sampling	61.41	59.30	75.45	59.34	50.74	51.45	68.66	51.35
w/o RC	61.64	60.28	75.65	60.64	50.92	51.31	68.26	51.65
w/o LCE	51.89	53.79	70.06	53.45	33.71	40.70	52.10	40.86
w/o CGGC	55.31	53.77	68.26	53.95	45.24	47.43	60.88	47.55

TABLE IV
COMPARED WITH DIFFERENT GRAPH NETWORKS FOR THE GLOBAL INFORMATION AGGREGATION BASED ON BERT_{base} ENCODER. BERT-BASE MEANS WITHOUT ANY GLOBAL GRAPH REASONING

Model	Closed				Open			
	AUC	F1	P@500	P@1000	AUC	F1	P@500	P@1000
BERT-base	55.37	52.19	67.65	54.84	45.20	48.08	62.26	47.05
BERT-GAT	58.46	56.89	73.05	56.65	47.12	49.41	66.87	50.15
BERT-GCN	60.17	59.37	73.25	59.54	49.74	50.64	67.86	51.05
BERT-GraphSAGE	60.93	60.06	74.85	60.94	49.88	51.28	67.07	51.45
BERT-LGCR (Ours)	**63.17**	**61.67**	**76.65**	**61.84**	**51.48**	**52.96**	**70.06**	**52.19**

tively and keep all other structures and hyperparameters unchanged. The results are shown in Table IV. From the experimental results we can see that compared with the widely used graph neural networks, our method has a significant improvement in performance (3.23% improvement in AUC and 2.66% improvements in F1 in the closed setting). This proves that it is effective to introduce causal effects into the neighborhood propagation of graph neural networks. In addition, it also shows that confusing filtering through causal effects is an effective approach to the global reasoning.

C. Further Analysis

1) Impacts of Important Hyperparameters: The three hyperparameters in (17) are critical to the performance of our model, so we use the grid search to find the most appropriate values under the closed setting. Firstly, we need to balance the weights between local and global reasoning, so we set $\alpha = 1.0$ and $\gamma = 0.01$ to search for β, the results are shown in Fig. 4. We can see that the model achieves the best performance when $\beta = 0.8$. Then, we need to determine how the regularization constraint exerts on the model. Therefore, we set $\alpha = 1.0$ and $\beta = 0.8$ to search for γ, the results are shown in Fig. 5. We can see that the best value for γ is 0.01. Finally, the hyperparameters we use for model training are $\alpha = 1.0$, $\beta = 0.8$ and $\gamma = 0.01$.

2) Computational Complexity Analysis: From the perspective of computation, our model can be mainly decomposed into three parts, which are BERT encoder, LCE and CGGC. Therefore, we will mainly analyze the three modules from two aspects, which are time complexity and space complexity. To this end, we gradually delete specific modules on LGCR similar to ablation experiments in Section IV-B-2), and perform statistics on the running time and the number of parameters of

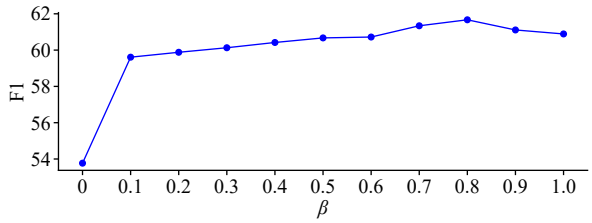

Fig. 4. The hyperparameter search results for β.

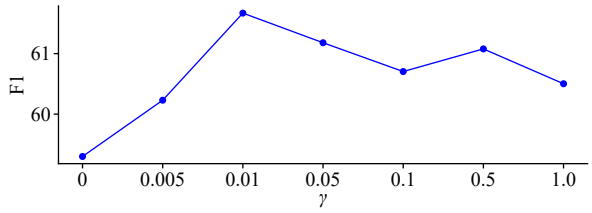

Fig. 5. The hyperparameter search results for γ.

different models. In addition to the three main modules of appeal, we also conduct the experiment on connection sampling. We use four RTX3090 GPUs to conduct experiments on the inference of validation set, and each GPU corresponds to one CPU core for model calculation and two CPU cores for data processing. The experimental results are shown in Table V, and "w/o CGGC, LEC" means that only the BERT encoder and a linear layer are left. From the results we can see that:

- In terms of the number of parameters, almost 90% of the parameters are concentrated on BERT. This is because BERT has the embedding layer for large-scale vocabulary and the multi-layer transformer structure. This is the foundation of its powerful representation capabilities. In addition, CGGC uses

TABLE V
THE RESULTS OF COMPUTATIONAL COMPLEXITY ANALYSIS, WHERE "W/O" INDICATES WITHOUT, "S" REPRESENTS SECOND, "IT/S" IS USED TO DESCRIBE HOW MANY INSTANCES ARE CALCULATED PER SECOND

	Closed		Open		Parameters
	Time	Speed	Time	Speed	
LGCR	112 s	49.71 it/s	195 s	28.55 it/s	119.14M
w/o sampling	114 s	48.84 it/s	198 s	28.12 it/s	119.14M
w/o LEC	104 s	53.54 it/s	185 s	30.10 it/s	119.14M
w/o CGGC	98 s	56.82 it/s	183 s	30.43 it/s	108.95M
w/o CGGC, LEC	89 s	62.56 it/s	171 s	32.56 it/s	108.95M

multiple networks to handle different types of edges, so it also has 10.19M parameters. Since the LCE and the local reasoning module share parameters, no additional parameters are added.

- After deleting the connection sampling module, the running time increases. The reason is that although the connection sampling module will increase the calculation amount, it can cut off the connection between irrelevant nodes to reduce the calculation of CGGC.
- Similarly to the proportion of parameters, in LGCR, the calculation of BERT occupies most of the time. In addition, due to the need to calculate the ITE and node representation of each semantic unit, the LCE (7.14% in the closed setting) and CGGC (12.50% in the closed setting) also consume a certain amount of time. Overall, the increased computational requirements of LCE and CGGC over BERT are acceptable relative to the performance improvement.

3) Comparisons of Local Reasoning Ability: The local causal effects are estimated based on the local reasoning module, so the better local reasoning ability can also estimate more accurate causal effects. To this end, we evaluate the local reasoning ability of different methods. Due to the excessive number of confusing paths, we only calculate the F1 and accuracy of valid paths. The results are shown in Table VI. We can see that the accuracy of local reasoning is low since the relation cannot determine accurately based on the single path in cross-document RE. This also shows that it is difficult to distinguish between relevant paths and irrelevant paths directly using classification algorithms. For LGCR, the local reasoning ability of BERT$_{base}$-based and RoBERTa$_{large}$-based models is not significantly different in the open setting, indicating that their accuracy has reached an upper bound, and it is of little significance to continue to increase model parameters in the real open domain scenarios. LGCR has a significant advantage over other methods in local reasoning because we reduce the influence of irrelevant information based on the local causal effect. While other methods will cause confusion. This demonstrates the effectiveness of our method.

4) Impacts of the Number of Confusing Paths: Since the reasoning paths are obtained by retrieval, they inevitably contain confusing paths irrelevant to the target. To evaluate the impact of such confusing paths, we calculate the cumulative F1 metric when the number of confusing paths goes from 0 to k for BERT-attn, BERT-GAIN, BERT-Ecrim and BERT-LGCR.

TABLE VI
COMPARISON OF LOCAL REASONING ABILITIES OF DIFFERENT METHODS. THE F1 ARE AVERAGED BY "WEIGHTED"

Model	Closed		Open	
	Acc	F1	Acc	F1
BERT-attn$_{base}$	25.80	33.32	18.12	25.03
CorefBERT-RE$_{base}$	28.38	36.56	17.93	25.12
GCN-BERT$_{base}$	35.57	42.63	19.33	26.12
PAR-DRE-BERT$_{base}$	38.08	46.73	20.16	28.32
Ecrim-BERT$_{base}$	41.64	48.96	12.42	18.90
LGCR-BERT$_{base}$	**44.84**	**53.27**	**21.29**	**30.10**
RoBERTa-attn$_{large}$	29.48	37.30	17.28	23.50
LGCR-RoBERTa$_{large}$	**47.11**	**56.76**	**21.90**	**30.82**

The results are shown in Figs. 6 and 7 for the closed and the open setting respectively. We can see that when the number of confusing paths is about $0 \sim 2$, the performance of different methods is similar, but when the number continues to increase, our method starts to show advantages. This demonstrates the effectiveness of our method in identifying confusing semantic units based on local causal effects, and the ability to better handle irrelevant information in global reasoning.

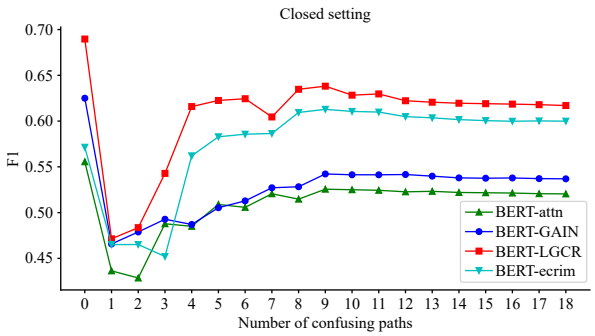

Fig. 6. The cumulative F1 when the number of confusing paths goes from 0 to k for the closed setting.

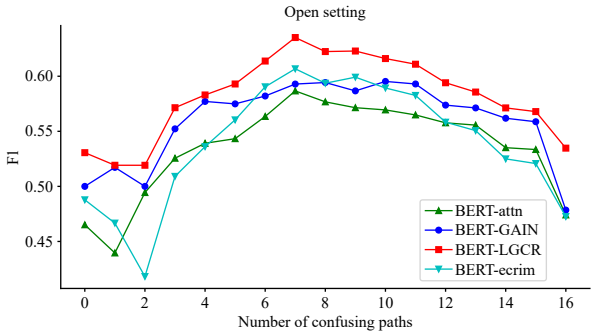

Fig. 7. The cumulative F1 when the number of confusing paths goes from 0 to k for the open setting.

We observe that in Fig. 6, there is a decrease in the performance curve when the number of confusing paths increases from 0 to 1. This is caused by the distribution of data. In the CodRED, since the reasoning paths are retrieved from Wikipedia, the number of reasoning paths associated with

each entity pair is not fixed. This is in line with the actual application scenario. To explain this phenomenon, we count the average number of reasoning paths corresponding to the different numbers of confusing paths in the Fig. 6. We find that when the number of confusing paths is 0, 1, 2 and 3, the corresponding average number of reasoning paths is 2.00, 1.26, 2.28 and 5.28, respectively. This shows that when the number of confusing paths is 1, the corresponding valid paths information is extremely sparse. With the increase of valid reasoning paths, the performance of the model gradually increases until the number of confusing paths is too large. In addition, a similar situation exists in the open setting since some entity pairs cannot retrieve enough 16 reasoning paths from Wikipedia. This phenomenon does not affect our conclusion that our method can better deal with irrelevant information in the confusing path. Besides, it also proves the premise that global reasoning of multiple paths is required to obtain the information in cross-document RE.

5) Evaluations of Local Causal Estimation Ability: Since our method uses LCE to guide the aggregation of global information, we evaluate the causal reasoning ability of LGCR in this section. Following the setting from [60]–[62], we use predictive accuracy as the evaluation metric. It refers to re-feeding the extracted results into the target model and measuring its fidelity by how it will recover the target predictions. Specifically, in our work, we obtain the local causal effect of the semantic unit corresponding to each node on the graph through LCE. On this basis, the local causal effects are sorted from small to large, and the feature expression of the corresponding semantic units is masked with a zero vector in order according to a proportion ρ. Then we re-input the masked features into the original model to observe the change of its accuracy. Besides, we provide two attribution methods [63] and the LCE guided by ground truth labels for comparison.

- **Gradient**, which calculates the gradient of the probability distribution of the final target output to the feature of semantic units, and uses it as the contribution of the feature to the result.
- **Attn**, which calculates the average attention weights of each node feature relative to the rest of the tokens based on the attention weights of the last layer of the BERT, and takes them as the importance of node features.
- **LGCR**, which is the method proposed in this paper that uses the local causal effects to measure the contribution of each node feature to the final prediction results.
- **LGCR-GT**, which is an extension of the LGCR, using the ground truth labels instead of the prediction results of the model to obtain the local causal effect. It can be considered as an upper limit of the local causal effect estimation.

The results are shown in Table VII for the closed setting and Table VIII for the open setting.

From the experimental results, we can see that our method outperforms the baseline in both the closed setting and the open setting. With the increase of masking ratio, the expression of features that are considered to be irrelevant to the target is inhibited, and the advantage of LGCR over Gradient and Attn becomes more obvious. When the masking ratio reaches to 70%, the F1 of LGCR in the closed setting only decreases by 13.30% (from 61.67 to 53.47), while the F1 of Gradient and Attn decrease by 17.48% (from 61.67 to 50.89) and 28.41% (from 61.67 to 44.15), respectively. Similarly, the F1 of LGCR in the open setting only decreases by 16.88% (from 52.96 to 44.02), while the F1 of Gradient and Attn decrease by 19.88% (from 52.96 to 42.43) and 25.98% (from 52.96 to 39.20), respectively. This shows that when the features with a low contribution to the target are masked, the features extracted by LGCR can better recover the results of the original model, while the similarity and gradient based methods are less effective. This also demonstrates the effectiveness of capturing key semantic units through LCE and using them to guide the global reasoning in cross-document RE.

LGCR-GT is an upper bound of LCE, which captures the local causal effects of different semantic units when our model can judge the relation between entities with complete accuracy. From the experimental results, we can see that with the increase of the masking ratio, the performance of the model continues to improve, even exceeding the results using the full node. The accuracy does not obviously decrease until the masking ratio reaches 90%, when the valid information is

TABLE VII
THE EVALUATION OF LOCAL CAUSAL ESTIMATION ABILITY FOR THE CLOSED SETTING. FULL NODE DENOTES THAT NO FEATURES OF SEMANTIC UNITS ARE MASKED

Mask ratio	Method	Closed			
		AUC	F1	P@500	P@1000
Full node	LGCR	63.17	61.67	76.65	61.84
$\rho = 30\%$	Gradient	58.13	58.17	71.06	58.34
	Attn	59.21	57.25	73.65	57.24
	LGCR	62.80	60.07	75.84	60.24
	LGCR-GT	69.23	64.05	83.23	64.04
$\rho = 40\%$	Gradient	57.03	57.00	70.06	57.04
	Attn	54.94	53.45	69.46	53.75
	LGCR	60.81	59.05	73.85	58.94
	LGCR-GT	69.58	64.10	82.64	64.54
$\rho = 50\%$	Gradient	54.70	55.56	68.66	55.45
	Attn	49.19	49.86	65.07	49.85
	LGCR	57.74	57.39	70.86	57.34
	LGCR-GT	69.04	64.05	82.24	64.24
$\rho = 60\%$	Gradient	52.22	53.82	66.67	54.15
	Attn	43.77	46.79	58.28	45.75
	LGCR	56.34	56.43	70.46	56.24
	LGCR-GT	69.37	64.33	84.23	64.94
$\rho = 70\%$	Gradient	47.66	50.89	63.67	50.85
	Attn	39.66	44.15	54.96	44.26
	LGCR	52.63	53.47	69.06	53.25
	LGCR-GT	67.81	63.58	84.23	64.04
$\rho = 90\%$	Gradient	27.39	35.66	44.71	34.97
	Attn	26.92	34.00	40.72	32.97
	LGCR	34.63	38.79	55.89	38.16
	LGCR-GT	50.21	50.78	72.06	50.15
$\rho = 95\%$	LGCR-GT	30.47	33.89	50.70	32.67

TABLE VIII
THE EVALUATION OF LOCAL CAUSAL ESTIMATION ABILITY FOR THE OPEN SETTING. FULL NODE DENOTES THAT NO FEATURES OF SEMANTIC UNITS ARE MASKED

Mask ratio	Method	Open			
		AUC	F1	P@500	P@1000
Full node	LGCR	51.48	52.96	70.06	52.19
ρ = 30%	Gradient	48.52	50.76	66.47	49.05
	Attn	49.55	50.53	65.67	49.65
	LGCR	51.45	51.18	68.26	50.75
	LGCR-GT	57.43	55.56	74.45	55.54
ρ = 40%	Gradient	47.03	48.26	64.07	48.55
	Attn	46.75	48.24	63.87	48.05
	LGCR	49.60	49.72	65.67	49.95
	LGCR-GT	57.63	55.56	74.65	55.74
ρ = 50%	Gradient	45.08	47.98	61.68	46.35
	Attn	41.93	44.35	58.88	44.36
	LGCR	47.30	48.26	62.87	47.75
	LGCR-GT	57.88	55.71	74.25	55.54
ρ = 60%	Gradient	42.21	45.24	59.88	45.55
	Attn	39.18	42.58	56.09	42.26
	LGCR	45.20	46.46	60.68	46.65
	LGCR-GT	57.64	55.19	74.25	54.75
ρ = 70%	Gradient	36.40	42.43	53.29	41.76
	Attn	34.50	39.20	48.30	38.46
	LGCR	42.53	44.02	56.89	43.96
	LGCR-GT	57.28	54.79	75.05	54.35
ρ = 90%	Gradient	21.93	31.79	34.93	29.97
	Attn	22.93	29.78	36.73	29.27
	LGCR	33.05	37.47	50.70	37.26
	LGCR-GT	51.05	50.52	71.46	50.55
ρ = 95%	LGCR-GT	34.80	37.43	55.09	36.96

masked. For the open setting, since we retrieve 16 reasoning paths for each sample, there are more semantic units. So this ratio raises to 95%. This proves our previous assumption that in cross-document RE, there are irrelevant confusing paths and semantic units, which need to be filtered. When the irrelevant confusing semantic information is accurately filtered, the performance of the model can be further improved. This also shows that in our method, LCE and RE have mutually reinforcing effects. That is, better LCE can enhance the performance the model, and more accurate RE can in turn enhance the LCE.

6) Case Studies: We show two case studies in Figs. 8 and 9 based on BERT$_{based}$ and RoBERTa$_{large}$, respectively. For each sample, the upper part of the figure is the reasoning path, where the head entity is represented in red, the tail entity is represented in blue, and the same background color represents the same bridging entity. We provide two reasoning paths for each sample, where the one with specific local relation is the relevant path, and the one with the local relation "N/A" is the irrelevant confusing path. The "[SEP]" is used to split the two documents in the reasoning path. The lower part of the figure is the global inference graph constructed based on these two reasoning paths, and the nodes in the graph correspond to the semantic units of the same number in the reasoning paths. For clearer visualization, we only present the mention nodes, omit some of the edges, and multiply the relative causal effect by ten. Beside the LCE proposed in this paper, we also use two other methods based on similarity to obtain the associations between semantic nodes, which are the similarity of cosine and the similarity of GAT. For the similarity of cosine, we calculate the similarity between each node feature before global graph reasoning by cosine distance. For the attention of GAT, we replace the CGGC in our model with GAT, and extract the attention weights between nodes from the last layer. Except for GAT, the weights of the edges presented are before softmax. The results are marked on the edges between nodes in the figure respectively.

From the degree of association between nodes, we can see the advantages of LCE over similarity-based methods. The most obvious difference lies in the degree of association between different nodes representing the same entity. For example, the head entity "Patty Bouvier" in Fig. 8 is represented by node "1" and node "6" in the two reasoning paths, respectively. Since these are different representations of the same entity, and the context is the interpretation of it, although BERT and GAT can encode context-related information in the entity representation, a high degree of similarity is still preserved between the two representations. This is also reflected in the figure. For both the similarity of cosine and the similarity of GAT, there is no significant difference in the degree of association between nodes "1" and nodes "6" compared with other edges. Based on this observation, we can see that the similarity-based methods are sub-optimal to tackle this kind of situation. In our method LGCR, we can see that the association between node "1" and "6" is very low. Therefore, in the global reasoning, the message propagation between the two nodes will be truncated, and the irrelevant confusing information in the reasoning path "2" will hardly affect the effective information in reasoning path "1". This is a significant advantage of our method over similarity-based methods. In addition, this association based on causal estimates also provides interpretability to our method.

V. CONCLUSION

In this paper, we propose a novel local-to-global causal reasoning (LGCR) network to solve the problem of aggregated reasoning in confusing information with similar features for cross-document RE. We propose a local causal estimation algorithm to innovatively use the local causal effects of semantic units to distinguish confusing information in retrieved documents from the open domain. For global reasoning, we propose a causality guided global reasoning algorithm to filter confusing information and control the message propagation in a global graph by calculating the relative causal relation between nodes through the local causal effects. The experimental results under the closed and the open settings demonstrate that our method effectively discriminates target-

Head entity: Patty Bouvier; **Tail entity**: Selma Bouvier; **Relation**: Sibling;

Local causal estimation

Reasoning path 1
... A list published in 2008 by the same organization included Smithers; Patty Bouvier [1], Marge Simpson [2]'s lesbian sister, was included on both lists. ... [SEP] ... She is best known for her voice role as Marge Simpson [3] on the animated television series "The Simpsons [4]". She also voices other characters for the show, including Marge's mother, Jacqueline Bouvier, and sisters Patty and Selma Bouvier [5]. ...
Local relation: sibling

Reasoning path 2
... Aside from a short - lived relationship with Patty Bouvier [6], most of Skinner's love life has focused on Edna Krabappel. ... [SEP] Troy McClure is a fictional character in the American animated sitcom "The Simpsons [7]". ... He appears as the main character in "A Fish Called Selma", in which he marries Selma Bouvier [8] to aid his failing career and quash rumors about his personal life. ... Following Hartman's murder at the hands of his wife in 1998, two of his "Simpsons [9]" characters were retired, with Hartman's final appearance as McClure occurring in the tenth season episode "Bart the Mother" four months later. ...
Local relation: N/A

Causality of CGGC	Similarity of cosine	Attention of GAT

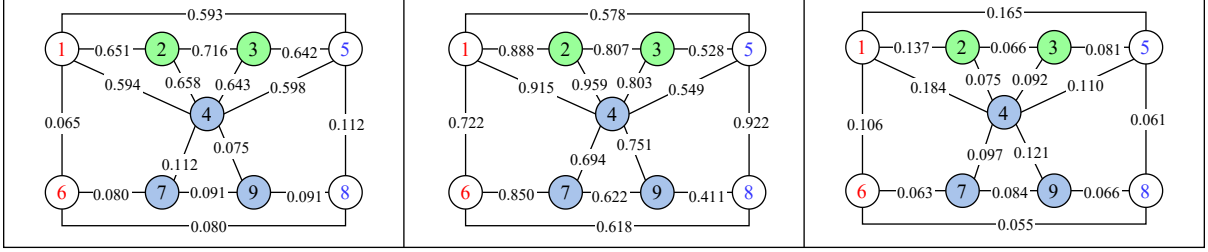

Fig. 8. The case study of our LGCR methods based on BERT$_{base}$. We also provide the similarity of cosine and the similarity of GAT for comparison.

Head entity: Alan fitz Walter; **Tail entity**: House of stuart; **Relation**: Family

Local causal estimation

Reasoning path 1
Alan fitz Walter [1], 2nd High Steward of Scotland [2]... Alan [3] was the son and heir of Walter fitz Alan [4]... [SEP] The House of Stuart [5], originally Stewart, was a Europen royal house of Scotland [6], with historical connections to Brittany. The family name itself comes from the office of High Steward of Scoland [7], which had been held by the family scion Walter fitz Alan [8]...
Local relation: Family

Reasoning path 2
...it could be evidence of an attempt by Raghnall - ... - to secure an alliance with Walter's son and successor, Alan fitz Walter [9], Steward of Scotland [10]... [SEP] ... The House of Stuart [11] in England had been involved in attempts to secure peace in Europe...
Local relation: N/A

Causality of CGGC	Similarity of cosine	Attention of GAT

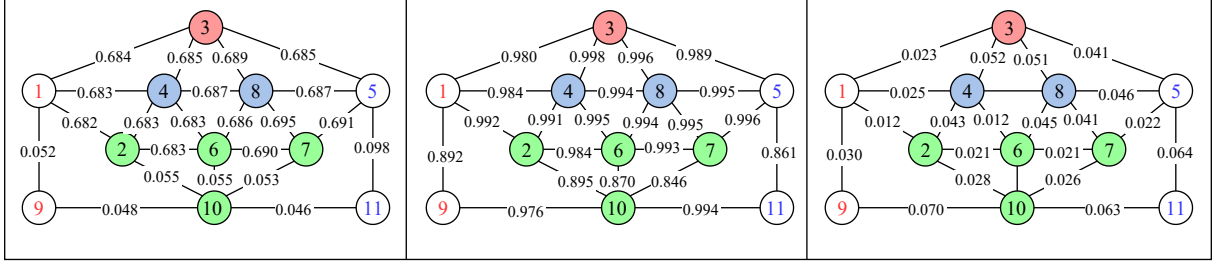

Fig. 9. The case study of our LGCR methods based on RoBERTa$_{large}$. We also provide the similarity of cosine and the similarity of GAT for comparison.

related and irrelevant information, and select effective information from confusing information for aggregated reasoning. In future work, we hope to apply this method of using causal reasoning to deal with confusing information to more fields, such as multimodal image-text joint reasoning.

REFERENCES

[1] B. Distiawan Trisedya, G. Weikum, J. Z. Qi, and R. Zhang, "Neural relation extraction for knowledge base enrichment," in *Proc. 57th Annu. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 229–240.

[2] M. Yu, W. P. Yin, K. S. Hasan, C. dos Santos, B. Xiang, and B. W. Zhou, "Improved neural relation detection for knowledge base question answering," in *Proc. 55th Annu. Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 2017, pp. 571–581.

[3] H. Z. Yu, H. S. Li, D. H. Mao, and Q. Cai, "A relationship extraction method for domain knowledge graph construction," *World Wide Web*, vol. 23, no. 2, pp. 735–753, Mar. 2020.

[4] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," in *Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, 2012, pp. 1201–1211.

[5] Y. K. Lin, S. Q. Shen, Z. Y. Liu, H. B. Luan, and M. S. Sun, "Neural relation extraction with selective attention over instances," in *Proc. 54th Annu. Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016, pp. 2124–2133.

[6] P. D. Qin, W. R. Xu, and W. Y. Wang, "Robust distant supervision

relation extraction via deep reinforcement learning," in *Proc. 56th Annu. Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018, pp. 2137–2147.

[7] J. Li, Y. P. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Y. Lu, "BioCreative V CDR task corpus: A resource for chemical disease relation extraction," *Database*, vol. 2016, pp. baw068, May 2016.

[8] Y. Yao, D. M. Ye, P. Li, X. Han, Y. K. Lin, Z. H. Liu, Z. Y. Liu, L. X. Huang, J. Zhou, and M. S. Sun, "DocRED: A large-scale document-level relation extraction dataset," in *Proc. 57th Annu. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 764–777.

[9] W. X. Zhou, K. Huang, T. Y. Ma, and J. Huang, "Document-level relation extraction with adaptive thresholding and localized context pooling," in *Proc. 35th AAAI Conf. Artificial Intelligence*, 2021, pp. 14612–14620.

[10] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, Oct. 2014.

[11] Y. Yao, J. J. Du, Y. K. Lin, P. Li, Z. Y. Liu, J. Zhou, and M. S. Sun, "CodRED: A cross-document relation extraction dataset for acquiring knowledge in the wild," in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2021, pp. 4452–4472.

[12] G. S. Nan, Z. J. Guo, I. Sekulic, and W. Lu, "Reasoning with latent structure refinement for document-level relation extraction," in *Proc. 58th Annu. Meeting of the Association for Computational Linguistics*, 2020, pp. 1546–1557.

[13] B. Li, W. Ye, Z. H. Sheng, R. Xie, X. Y. Xi, and S. K. Zhang, "Graph enhanced dual attention network for document-level relation extraction," in *Proc. 28th Int. Conf. Computational Linguistics*, 2020, pp. 1551–1560.

[14] S. Zeng, R. X. Xu, B. B. Chang, and L. Li, "Double graph based reasoning for document-level relation extraction," in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2020, pp. 1630–1640.

[15] W. Xu, K. H. Chen, and T. J. Zhao, "Document-level relation extraction with reconstruction," in *Proc. 35th AAAI Conf. Artificial Intelligence*, 2021, pp. 14167–14175.

[16] H. R. Wu, W. Chen, S. Xu, and B. Xu, "Counterfactual supporting facts extraction for explainable medical record based diagnosis with graph network," in *Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1942–1955.

[17] Q. T. Wu, H. R. Zhang, J. C. Yan, and D. Wipf, "Handling distribution shifts on graphs: An invariance perspective," in *Proc. 10th Int. Conf. Learning Representations*, 2021.

[18] C. Peng and S. Athey, "Stable learning establishes some common ground between causal inference and machine learning," *Nat. Mach. Intell.*, vol. 4, no. 2, pp. 110–115, Feb. 2022.

[19] J. Pearl, *Causality*. 2nd ed. Cambridge, UK: Cambridge University Press, 2009.

[20] D. J. Zeng, K. Liu, S. W. Lai, G. Y. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proc. 25th Int. Conf. Computational Linguistics: Tech. Paper*, Dublin, Ireland, 2014, pp. 2335–2344.

[21] L. L. Wang, Z. Cao, G. de Melo, and Z. Y. Liu, "Relation classification via multi-level attention CNNs," in *Proc. 54th Annu. Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016, pp. 1298–1307.

[22] J. Feng, M. l. Huang, L. Zhao, Y. Yang, and X. Y. Zhu, "Reinforcement learning for relation classification from noisy data," in *Proc. 32nd AAAI Conf. Artificial Intelligence*, New Orleans, USA, 2018, pp. 5779–5786.

[23] N. Y. Zhang, S. M. Deng, Z. L. Sun, J. Y. Chen, W. Zhang, and H. J. Chen, "Relation adversarial network for low resource knowledge graph completion," in *Proc. Web Conf.*, Taipei, China, 2020, pp. 1–12.

[24] H. Wang, C. Focke, R. Sylvester, N. Mishra, and W. Wang, "Fine-tune Bert for docRED with two-step process," arXiv preprint arXiv: 1909.11898, 2019.

[25] H. Z. Tang, Y. N. Cao, Z. Y. Zhang, J. X. Cao, F. Fang, S. Wang, and P. F. Yin, "HIN: Hierarchical inference network for document-level relation extraction," in *Proc. 24th Pacific-Asia Conf. Knowledge Discovery and Data Mining*, Singapore, 2020, pp. 197–209.

[26] S. Zeng, Y. T. Wu, and B. B. Chang, "SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction," in *Proc. Findings of the Association for Computational Linguistics*, 2021, pp. 524–534.

[27] Q. Z. Huang, S. Q. Zhu, Y. S. Feng, Y. Ye, Y. X. Lai, and D. Y. Zhao, "Three sentences are all you need: Local path enhanced document relation extraction," in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. Natural Language Processing*, 2021, pp. 998–1004.

[28] Z. Q. Wang, S. C. Gao, M. C. Zhou, S. Sato, J. J. Cheng, and J. H. Wang, "Information-theory-based nondominated sorting ant colony optimization for multiobjective feature selection in classification," *IEEE Trans. Cybern.*, 2022, doi: 10.1109/TCYB.2022.3185554.

[29] S. Ramírez-Gallego, H. Mouriño-Talín, D. Martínez-Rego, V. Bolón-Canedo, J. M. Benítez, A. Alonso-Betanzos, and F. Herrera, "An information theory-based feature selection framework for big data under apache spark," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 48, no. 9, pp. 1441–1453, Sept. 2018.

[30] R. Q. Lu, X. L. Jin, S. M. Zhang, M. K. Qiu, and X. D. Wu, "A study on big knowledge and its engineering issues," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 9, pp. 1630–1644, Sept. 2019.

[31] S. Athey, "Beyond prediction: Using big data for policy problems," *Science*, vol. 355, no. 6324, pp. 483–485, Feb. 2017.

[32] D. Wu, Y. He, X. Luo, and M. C. Zhou, "A latent factor analysis-based approach to online sparse streaming feature selection," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 52, no. 11, pp. 6744–6758, Nov. 2022.

[33] H. Wu, X. Luo, and M. C. Zhou, "Advancing non-negative latent factorization of tensors with diversified regularization schemes," *IEEE Trans. Serv. Comput.*, vol. 15, no. 3, pp. 1334–1344, Jun. 2022.

[34] D. Wu, M. S. Shang, X. Luo, and Z. D. Wang, "An L_1-and-L_2-norm-oriented latent factor model for recommender systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5775–5788, Oct. 2022.

[35] L. Hu, S. C. Yang, X. Luo, and M. C. Zhou, "An algorithm of inductively identifying clusters from attributed graphs," *IEEE Trans. Big Data*, vol. 8, no. 2, pp. 523–534, Apr. 2022.

[36] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Machine Learning*, Sydney, Australia, 2017, pp. 3145–3153.

[37] R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, Apr. 1983.

[38] Z. Y. Shen, P. Cui, K. Kuang, B. Li, and P. X. Chen, "Causally regularized learning with agnostic data selection bias," in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, Republic of Korea, 2018, pp. 411–419.

[39] K. Kuang, P. Cui, S. Athey, R. X. Xiong, and B. Li, "Stable prediction across unknown environments," in *Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, London, UK, 2018, pp. 1617–1626.

[40] D. Xu, C. W. Ruan, E. Korpeoglu, S. Kumar, and K. Achan, "Adversarial counterfactual learning and evaluation for recommender system," in *Proc. 34th Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 13515–13526.

[41] M. Oberst and D. Sontag, "Counterfactual off-policy evaluation with gumbel-max structural causal models," in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, USA, 2019, pp. 4881–4890.

[42] D. Kaushik, E. H. Hovy, and Z. C. Lipton, "Learning the difference that makes a difference with counterfactually-augmented data," in *Proc. 8th Int. Conf. Learning Representations*, Addis Ababa, Ethiopia, 2019.

[43] T.-J. Fu, X. E. Wang, M. F. Peterson, S. T. Grafton, M. P. Eckstein, and W. Y. Wang, "Counterfactual vision-and-language navigation via adversarial path sampler," in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 71–86.

[44] Q. F. Zhu, W.-N. Zhang, T. Liu, and W. Y. Wang, "Counterfactual off-policy training for neural dialogue generation," in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2020, pp. 3438–3448.

[45] J. C. Weiss, F. Kuusisto, K. Boyd, J. Liu, and D. Page, "Machine learning for treatment assignment: Improving individualized risk attribution," in *Proc. AMIA*, San Francisco, USA, 2015, pp. 1306.

[46] T. Zhao, G. Liu, D. H. Wang, W. H. Yu, and M. Jiang, "Learning from

counterfactual links for link prediction," in *Proc. 39th Int. Conf. Machine Learning*, Baltimore, USA, 2022, pp. 26911–26926.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, USA, 2017, pp. 6000–6010.

[49] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *Proc. 5th Int. Conf. Learning Representations*, Toulon, France, 2017.

[50] E. Jang, S. X. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *Proc. 5th Int. Conf. Learning Representations*, Toulon, France, 2017.

[51] E. J. Gumbel, *Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures*. Washington, USA: US Government Printing Office, 1954.

[52] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. 32nd Int. Conf. Neural Information Processing Systems*, Red Hook, USA, 2018. pp. 7786–7795.

[53] Y. F. Sun, C. M. Cheng, Y. H. Zhang, C. Zhang, L. Zheng, Z. D. Wang, and Y. C. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, USA, 2020, pp. 6397–6406.

[54] D. M. Ye, Y. K. Lin, J. J. Du, Z. H. Liu, P. Li, M. S. Sun, and Z. Y. Liu, "Coreferential reasoning learning for language representation," in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2020, pp. 7170–7186.

[55] G. D. Duan, Y. R. Dong, J. Y. Miao, and T. X. Huang, "Position-aware attention mechanism-based bi-graph for dialogue relation extraction," *Cognit. Comput.*, vol. 15, no. 1, pp. 359–372, Jan. 2023.

[56] F. Q. Wang, F. Li, H. Fei, J. Y. Li, S. Q. Wu, F. F. Su, W. X. Shi, D. H. Ji, and B. Cai, "Entity-centered cross-document relation extraction," in *Proc. Conf. Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022, pp. 9871–9881.

[57] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learning Representations*, Toulon, France, 2017.

[58] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. 6th Int. Conf. Learning Representations*, Vancouver, Canada, 2018.

[59] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, USA, 2017, pp. 1025–1035.

[60] J. B. Chen, L. Song, M. Wainwright, and M. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *Proc. 35th Int. Conf. Machine Learning*. Stockholm, Sweden, 2018, pp. 883–892.

[61] J. Liang, B. Bai, Y. R. Cao, K. Bai, and F. Wang, "Adversarial infidelity learning for model interpretation," in *Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, 2020, pp. 286–296.

[62] X. Wang, Y.-X. Wu, A. Zhang, X. N. He, and T.-S. Chua, "Towards multi-grained explainability for graph neural networks," in *Proc. 35th Conf. Neural Information Processing Systems*, 2021, pp. 18446–18458.

[63] S. Jain, S. Wiegreffe, Y. Pinter, and B. C. Wallace, "Learning to faithfully rationalize by construction," in *Proc. 58th Annu. Meeting of the Association for Computational Linguistics*, 2020, pp. 4459–4473.

Haoran Wu received the B.E. degree in automation major from North China Electric Power University in 2018. He is currently the doctoral students in the Pattern Recognition and Intelligent System at Institute of Automation, Chinese Academy of Sciences, advised by Prof. Bo Xu. His main research interests include natural language processing, cognitive reasoning and human-AI hybrid intelligence.

Xiuyi Chen received the Ph.D. degree (2022) in pattern recognition and intelligent system from Institute of Automation, Chinese Academy of Sciences, advised by Prof. Bo Xu. Previously, he received the B.Sc. degree (2017) in Department of Control Science and Engineering from Jilin University. His current interests include dialogue system, multimodal learning and speech & language.

Zefa Hu received the B.S. degree in automation from Huazhong Agricultural University in 2018. He is currently a Ph.D. candidate at School of Artificial Intelligence, University of Chinese Academy of Sciences, and studying at Institute of Automation, Chinese Academy of Sciences.

Jin Shi is a Research Assistant in the Institute of Automation, Chinese Academy of Sciences, where he received the Ph.D. degree (2021) in the major of pattern recognition and intelligent system, advised by Prof. Bo Xu. Previously, he received the B.Sc. degree (2012) in School of Instrumentation and Optoelectronic Engineering from Beihang University. His current interests include cross-modal modeling, multimodal learning, dialogue system, speech recognition and speech separation.

Shuang Xu is a Professor in Institute of Automation, Chinese Academy of Science. Her main research interests include natural language processing and understanding, human-AI hybrid intelligence.

Bo Xu is a Professor, the Director of the Institute of Automation Chinese Academy of Sciences, and also Deputy Director of the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences. His main research interests include brain-inspired intelligence, brain-inspired cognitive models, natural language processing and understanding, brain-inspired robotics.