ERDAP: A Novel Method of Event Relation Data Augmentation Based on Relation Prediction

Ruijuan Hu
Department of Data and Target, Information Engineering University, China
huruijuan01@126.com

Haiyan Liu
Department of Data and Target, Information Engineering University, China
Liuharry2020@163.com

Wentao Yu
Department of Data and Target, Information Engineering University, China
wenteyu@163.com

Abstract: *Event relation extraction is a key aspect in the fields of event evolutionary graph construction, knowledge question and answer, and intelligence analysis, etc., Currently, supervised learning methods that rely on large amounts of labeled data are mostly used; however, the size of existing event relation datasets is small and cannot provide sufficient training data for the models. To alleviate this challenging research question, this study proposes a novel data augmentation model, called Event Relation Data Augmentation based on relationship Prediction (ERDAP), that allows both semantic and structural features to be taken into account without changing the semantic relation label compatibility, uses event relation graph convolutional neural networks to predict event relations, and expands the generated high-quality event relation triples as new training data for the event relation texts. Experimental evaluation using event causality extraction method on Chinese emergent event dataset shows that our model significantly outperforms existing text augmentation methods and achieves desirable performance, which provides a new idea for event relation data augmentation.*

Keywords: *Relation prediction, event relation, data augmentation, graph convolutional networks, causality extraction.*

Received October 12, 2022; accepted June 19, 2023
https://doi.org/10.34028/iajit/21/1/6

1. Introduction

Events contain a large number of internal constituent structures (e.g., participants, time, place, etc.) and external associations (e.g., semantic relations such as causality, co-reference, temporal order, etc.,). The relation extraction for texts containing a large number of events can achieve a deeper understanding of the text, and event relation extraction is an important semantic processing task in the field of natural language processing [6]. As shown in Figure 1, E1, E2, and E3 are three event texts, and the occurrence of event E1 causes events E2 and E3 to occur, and there is an obvious causal relation between the event pairs (E1, E2) and (E1, E3), and usually, there is often a sequential relation in the temporal order between the events with causal relations.

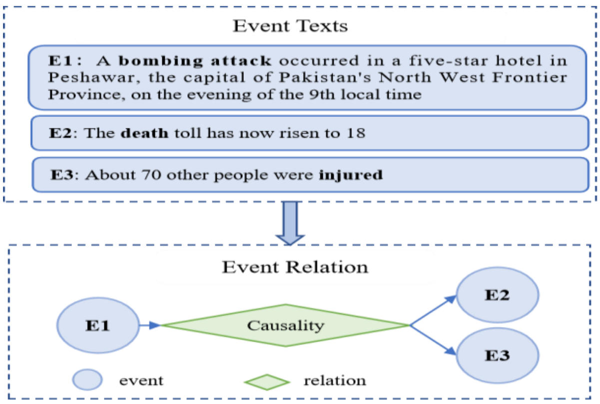

Figure 1. Illustrations of event relations.

Currently, there is no unified framework for defining event relation in either cognitive or linguistic science, resulting in generally small datasets for event relation extraction, which poses a challenge for adequate training of models, a problem that is particularly acute in intricate event causality extraction. Therefore, data augmentation methods are needed to synthesize new data from existing training data to deal with data scarcity in order to improve the performance of downstream models.

Data augmentation methods have become a key factor in the performance improvement of various neural network models, mainly in the fields of computer vision and speech recognition, such as cropping, padding, flipping and shifting along temporal and spatial dimensions by means of transformations [8, 10], however, for textual data, such transformations are usually ineffective or distort the text, resulting in grammatical and semantic incorrectness. It makes data augmentation more challenging in the textual domain, especially event relational texts. Data augmentation for text usually include using back translation [21], Easy Data Augmentation simple data augmentation (EDA) [18] and contextual augmentation using language models [9, 22], which consider more word-level features and do not consider deeper semantic and structural features.

This study proposes a novel model for event relation data augmentation, by utilizing a hierarchical method to predict event relation. In detail, we construct event relation graph to input the Encoder side. Next, we use an end-to-end approach to encode events to generate potential hidden feature vector representations of target events. Finally, at the Decoder side, we perform event relation prediction filter based on the prediction results, and generate new training data to achieve the data. The effect of data augmentation is achieved.

Our major contributions of our work can be summarized as follows:

1. Ease of use when formatting individual papers, we propose an Event Relation Data Augmentation based on relation Prediction (ERDAP**)** model capable of augmenting event-relational sentences without changing the compatibility of semantic relational labels. So far as we know this is the first one of that uses relation prediction for data augmentation.
2. The novelty of ERDAP is that an EventRGCN is used as the event relation prediction model to predict some high-quality new event relation samples.
3. Our ERDAP model achieves better performances than the previous models and provides a new idea for event relation data augmentation.

2. Related Work

Regarding datasets construction for event relation extraction, Caselli and Vossen [3] annotated EventStoryLine Corpus (ESC) consisting of 258 documents, 4316 sentences and 1770 event pairs. Mirza and Tonelli [15] annotated event causal relations and released a corpus called Causal-TimeBank containing 184 documents, 6813 events, and 318 event pairs. Chinese Emergency Corpus (CEC) contained 332 documents and 3069 event pairs. These small-scale data cannot provide sufficient data for the event relation extraction model to train and test. Therefore, it is necessary to generate a large amount of annotated data from a small amount of annotated data under the premise of keeping the label semantics unchanged as far as possible, namely text augmentation. Zuo *et al*. [25] proposed a learnable knowledge guided data augmentation method for event causality identification tasks. By introducing external lexical knowledge and common sense knowledge, the dual learning mechanism was used to learn how to generate new data related to ECI tasks.

Text augmentation has attracted a lot of interests among researchers and three types of methods are proposed. The first category is back translation which is simple but depends on the ability of the machine translation model. Yu *et al*. [21] translate the original text into another language and back again using a machine translation model, or translate from a to b to c and then back again. The second was EDA that is achieved by performing synonym replacement, random insertion, random swap, and random deletion of data using words in the prepared lexicology. It is a text enhancement scheme which is suitable for small amount of data and has no extra resource overhead. However, EDA does not take the context into account. The third category is contextual augmentation in which text replacement by using language models such as BERT [5] instead of fixed dictionaries can avoid the disadvantages of limited dictionaries and improve the correct rate of synonyms in different contexts, thus improving data quality [19]. With the emergence of GPT, GPT-2, GPT-3 and other models with amazing effects on text generation task, the method of text augmentation using language generation model has emerged [7, 11, 14]. The very recent work Language-Model-Based Data Augmentation (LAMBDA) utilize Generative Pre-trained Transformer (GPT) for text augmentation [1] to ensure that the newly generated data sets have similar distribution to some small data sets. However, the language model is only a shallow connection between the forward and backward models, and its predictive ability is limited to a short range. Too many substitutions may affect the sentence semantics.

3. Approach

Event relational data not only requires that two events in the text must have a certain probability of correlation, but also needs to rely on the actual requirements of text expression. We subtly regard event relation data augmentation as a relation prediction task, and propose an ERDAP. An improved graph convolutional neural network model is used as the event relation prediction model taking into account the contextual semantic information to help us predict some new event relation samples with high quality, so as to achieve the effect of data augmentation. Figure 2 schematically visualizes out approach.

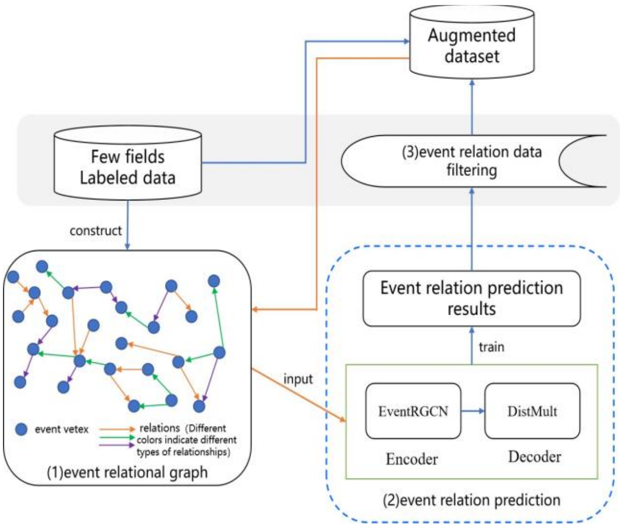

Figure 2. Event relational data augmentation framework.

The framework consists of three major components:

1. Event relational graph, which construct event relational graph from a small amount of existing domain annotation data, and the input to the

EventRGCN (3.1.).
2. Relation prediction, which adopts EventRGCN model as an encoder and DistMult model as a decoder to train and generate relation prediction results (3.2.).
3. Data filtering, which filters and selects the reasonable event relation data as augmented data and fuses with the original training set to generate an expanded samples, and the selected. Finally, the augmented data is a set as augmented dataset (3.3.).

3.1. Event Relational Graph

The structure of a relational graph network with n events can be represented by a directed graph $G=(V, E, R)$. Each vertex in the graph represents an event which is an event text containing trigger word and argument role, any vertex is denoted as $v_i \in V$, the edge between the vertex v_i and $v_j (i, j \in \{1, 2, \ldots, n\})$ and is denoted as $(v_i, v_j) \in E$, the edge represents the event relation type, using $r_k \in R (k \in \{1, 2, \ldots, m\})$. Therefore, the event relation text can be represented by the event triplet (event i, relation k, event j), denoted as (v_i, r_k, v_j). In Figure 2, blue circles denoting vertices represent events and arrows of different colors denoting directed edges represent different types of relations. The original event text is represented as event triples and the event relational graph is constructed.

3.2. Relation Prediction

The constructed event relational graph is a heterogeneous graph with multiple relations. The update of a vertex is determined by the vertices connected by different types of edges, and the edges of the same type are also oriented. Drawing on the design idea of the relational graph convolutional neural network model proposed by Schlichtkrull *et al.* [16], we propose an Event Relational data with Graph Convolutional Neural network (EventRGCN) model for event relation prediction. Graph Convolutional Network (GCN) is used to deal with the influence of different edge relations on event nodes in the graph, and more event relations are predicted through the local domain structured relation information in the event relational graph.

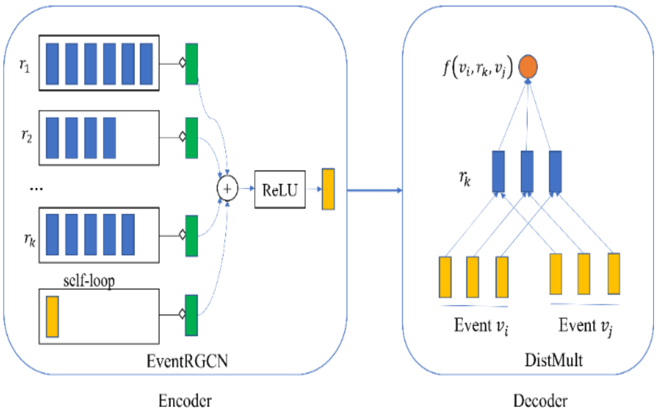

Figure 3. Relation prediction model containing an encoder EventRGCN and a decoder DistMult.

The pseudo-code for the event relation data augmentation Algorithm (1) based on relationship prediction is as follows:

Algorithm 1: Event Relation Data Augmentation Based on Relationship Prediction.

Input : Adjacencies, adjacency matrixes of Event relational graph G=(V, E, R). The dim of adjacencies is [node_num, node_num, adj_num], where node_num denotes the number of vertices in the graph, adj_num denotes the number of adjacencies.
Output : The new event relation data representing as data_enhance[].
1: pos_tripelts,rel_ent_freq,pos_rdf_num <- read(data:[(v_i, r_k, v_j)]) #Get positive samples, event-relation- frequency, number of positive samples
2: neg_triplets <- negativeSampling(pos_tripelts, part_pos_num, rel_ent_freq, pos_rdf_num, related_neg_num) # Generate a negative sample, part_pos_num is the number of randomly selected positive samples, set to 10. Specify the number of negative samples corresponding to a positive sample, related_neg_num is set to 30
3: inputs_triplets = combinePosNeg(pos_tripelts, neg_triplets, part_pos_num, related_neg_num) # fusion positive and negative samples
4: h_layer = Encoder(embedding_dim, adj_num) # the resutls of node embeddings, default = 200
5: dropout = Dropout(dropout_rate) # Proportion of vacant neurons, default = 0.1
6: decoder_inputs <- [vi_embs, vj_embs, rel_ids] #vi_embs, vj_embs, rel_ids respond to v_i, r_k, v_j, of (v_i, r_k, v_j)
7: score(decoder_inputs) <- Decoder(h_layer, edge_count,batch_size) # edge_count denotes number of edges, batch_size indicates the number of event relationship triples input at one time,default=100
8: L<- Cross-entropy()
9: outputs,data_enhance=[]
10: for i in range(len(inputs_triplets))
11: if 1 in outputs[k]:
12: rank = k # Data enhancement section: filtering negative samples with scores ranked in front of positive samples
13: if outputs[k][3] != 1:
14: data_enhance.append([outputs[k][0], outputs[k][1], outputs[k][2]])
15: else:
16: break
17: return data_enhance #output the new event relation data.

1. Encoder

EventRGCN model. The encoder is an EventRGCN that generates an implicit feature representation of the event node, as shown on the left side of Figure 3, and its main contribution is to encode the vertices of the event relational graph (blue part of the figure) and transform the vertices into feature vectors (embedding vector, green part of the figure), so that its input is the relation information in the local domain of the target event node, such as the relation type, the relation direction, and self-loop information of the node (self-loop, yellow part of the figure), etc., The output is the potential feature vector representation of the target event node (yellow part after the activation function).

In modeling the event text, considering that different event relations have different characteristics, for example, the temporal relation indicates the sequence of events in time, and the causal relation indicates the action relation between events, where the former event causes the latter event to occur. Therefore, different event relations are divided, and a graph structure is constructed for each relation separately, and the triple of event relations is transformed into multiple directed graphs as the encoder input. The nodes in the graph are updated as in Equation (1):

$$h_{v_i}^{l+1} = \sigma(\sum_{r \in R} \sum_{m \in N_{v_i}^r} \frac{1}{c_{v_i,r}} W_r^l h_{v_j}^l + W_0^l h_{v_i}^l) \quad (1)$$

$N_{v_i}^r$ denotes the set of neighboring nodes of a node v_i under the relation r. $c_{vi,r}$ is a normalized constant that takes the value of $|N_{v_i}^r|$. W_r is a linear transformation function that transforms the neighbor nodes of the same type of edge using a parameter matrix W_r^l, the number of W_r^l which is the number of edge types. σ is the activation function that expresses the summation of the multiplication including self-loops of the graph structure with the parameter matrix for each type of relation when propagating forward from the previous layer to the next layer. Here only one variable matrix is used as a trainable parameter for the input adjacency matrix. We use the adjacency matrix to represent the event relational graph.

The dimensions of the input adjacency matrix are, [node_num, node_num, adj_num], where node_num denotes the number of vertices in the graph and adj_num denotes the number of adjacency matrices. Suppose there are n types of relations in the data, and since the topological relations of outgoing and incoming edges are different, together with the part of self-loop (a unit matrix), then there are (2*n+1) adjacency matrices that need to be passed into the model. Before feeding into the model, the adjacency matrix needs to be normalized. The parameter matrix is expressed as a matrix of [adj_num, node_num, node_embedding_dim], where node_embedding_dim is the vertex encoding, and the feature representation after passing through the encoding is a matrix of [node_num, node_embedding_dim].

2. Decoder

The decoder is a tensor factor decomposition model that uses the feature vector representation of the encoder stage to predict labeled edges, using a scoring function that scores multiple positive and negative samples simultaneously and then evaluates the position of the positive samples in the ranking of all sample scoring results, and the model is considered effective if most of the positive samples are ranked relatively high. We use DistMult as the scoring function which is a semantic matching model proposed by Yang *et al*. [20] that performs well on relation prediction tasks, as shown on the right side of Figure 3. It uses a similarity-based scoring function to measure the plausibility of the facts. Thus, the input is a triple of positive and negative samples of data, i.e., multiple (v_i, r_k, v_j), where v_i and v_j denotes event i and event j, r_k is the relation between two events, and the scoring function is Equation (2):

$$f(v_i, r_k, v_j) = e_{v_i}^T R_{r_k} e_{v_j} \quad (2)$$

e_{vi}, e_{vj} are vector representations of head vertices and tail vertices in a triple, and each relation r_k corresponds to a diagonal matrix Rr. We expect to train representations of events and relations that result in higher scores for legitimate triples, and conversely, scores for unreasonably negative samples should be relatively low.

- Positive and negative sample selection: for the samples that are already in the event relational graph, they are called positive samples; one event is randomly selected from all events and replaced with v_i or v_j in a triple (here also randomly selected) as negative samples to simulate the model boundary, so that the model can pay more attention to this confusing part during training, and achieve the purpose of improving model performance and enhancing model robustness.
- Cross-entropy loss function: the training method of negative sampling is used, and for the observed samples, w negative samples are considered and optimized using the cross-entropy loss, namely in Equation (3):

$$\mathcal{L} = -\frac{1}{(1+w)|\hat{\mathcal{E}}|} \sum_{(v_i,r_k,v_j,y)\in T} y\, log\sigma\left(f(v_i, r_k, v_j)\right) + (1-y) log(1-\sigma(f(v_i, r_k, v_j))) \quad (3)$$

For each positive sample (v_i, r_k, v_j)$\in\hat{\mathcal{E}}$, generate w negative samples by randomly destroying v_i or v_j. In Equation (3), T is the set of positive samples. If (v_i, r_k, v_j) is a positive sample, then y=1, if it is a negative sample, then y=0. *()* is a sigmoid function. $ylog\sigma(f(v_i, r_k, v_j)$ is used to optimize the discrimination of positive samples and $(1\text{-}y)log(1\text{-}\sigma(f(v_i, r_k, v_j))$ is used to optimize the discrimination of negative samples.

3.3. Data Filtering

The scores of all triples can be obtained for the input event relational graph after the training of the relation prediction module. Since our aim is to generate more event relation data, the specific score of the final scoring is not important, the key is that the positive samples should be ranked as high as possible, and the effect of the model can be considered good if this condition is satisfied. The negative samples with higher scoring results than the positive samples are filtered out and used as the new triple event relation data.

4. Experiment

In order to verify the effectiveness of ERDAP data augmentation, various experiments have been conducted separately: relationship prediction, using different event relationship extraction models to experimentally validate and compare the effects before and after data augmentation, and using the same event relationship extraction model to compare the effects of different data augmentation methods.

4.1. Evaluation Dataset

- **Dataset***:* our proposed method will be evaluated on the CEC [23] and the constructed Chinese event relationship dataset Commission for Energy Regulation (CER). CEC contains five types of emergencies: earthquake, fire, traffic accident, terrorist attack, and food poisoning, with a total of 332 articles and 6013 events. 6 types of event relations with a total of 3069 pairs are causally related: concurrency, 9 pairs, composition, 12 pairs, causality, 949 pairs, accompaniment, 578 pairs, following, 794 pairs, and thought content, 727 pairs. The CER dataset is reconstructed based on the multi-domain causality dataset provided by liu (https://github.com/liuhuanyong/CausalDataset), with 1550 pairs of three types of event relationships: causality, 835 pairs, accompaniment, 390 pairs, and following, 325 pairs.

In order to verify the effectiveness of data augmentation, we conducted experimental validation and effectiveness comparison before and after data augmentation using different event relation extraction models and different data augmentation methods using the same event relation extraction model, respectively. For evaluation, we adopt precision, recall and F1-score (F1) as evaluation metrics, same as previous event relation extraction methods to ensure comparability.

4.2. Experiment Setting

In our implementations, the main hardware environment of this experiment is as follows: Inter Core i9-11900K CPU, 64GB physical memory, 1 NVIDIA GeForce RTX 3090 graphics card; the main software environment is as follows: Windows 10 64-bit OS, Python 3.6, TensorFlow 2.5. Event node embedding is set to 200 dimensions; the number of adjacency matrix passed into the encoder is 13 because there are 6 kinds of event relations; the number of event triples input at one time in the decoder, batch_size is set to 100, and dropout_rate is set to 0.1; the ratio of positive and negative samples is set to 1:10 because there are only 3069 pairs of event relations. A hyperparameter of EventRGCN model, num_bases, refers to the number of samplers, similar to the multi-convolutional kernel sampling in convolutional neural networks, which is set to -1 in this paper, that is, there is no need to consider the case of multi-sampling.

4.3. Main Results of Relation Prediction

For relation prediction, the Mean Reciprocal Rank (MRR) and HITS@n (Hits at n) methods are usually used for evaluation, where MRR represents the average of the inverse of the rank of all positive samples in their respective scoring results, and the higher the rank of positive samples, the higher the MRR score. HITS@n denotes the percentage of the top n scores of the positive sample among all sample results. The purpose of this paper is data augmentation, and the relation prediction module can predict the scores of positive samples, so it is natural that the event relation triple with higher scores than positive samples is reasonable. Based on this consideration, we are concerned with the ranking of positive samples and all negative samples ranked before positive samples. Therefore, we omitted the evaluation of MRR and HIT@n, and scored all samples directly, and sorted these scores in ascending order, with all negative samples ranked before the positive samples as the new event relation data.

We found that some duplicate negative samples were generated during the experiment, and analyzing the reasons, there are mainly two cases:

1) The replaced event is ranked higher.
2) The modified new triple happens to be the triple that exists in the event relational graph (the event that is not modified in the modified triple is already related to the replaced event).

Of course, we can also replace events by completely random sampling, so that the difference between a logical triple and a random triple can be seen more clearly in the scoring results.

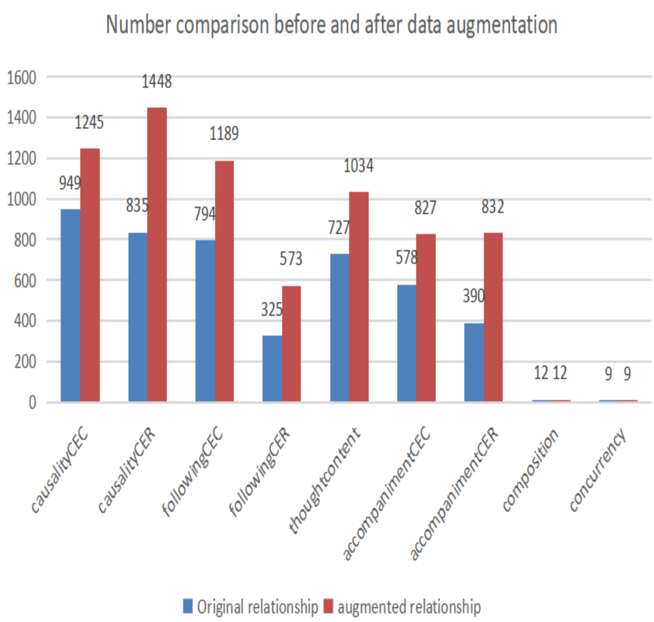

Figure 4. Data changes in the event relation triples before and after data augmentation.

Using the data augmentation method proposed in this paper, 10 iterations were performed to generate a total of 1248 and 1303 new event relation triples (after de-duplication) on the CEC and CER datasets, and the changes in the number of event relations are shown in Figure 4. It can be seen that the causality, following, thought content, and accompaniment relations increased by 296, 395, 307, and 249, respectively on the CEC, while the composition and concurrency relations, which originally had very small amounts of data, did not increase. On the CER, the causality, following, and accompaniment relationships increased by 613, 248, and 442, respectively.

Examples of the new event relation triples generated are shown in Table 1. From the text description, these relations are basically in accordance with the grammatical rules and contextual semantics, but because feature extraction is performed in the encoding stage without the support of common sense and background knowledge, relation data that do not conform to cognitive logic like sentence 5 (an earthquake in Chile causes tremors to be felt in Shanxi) can occur.

Table 1. Examples of generated event relation triples.

	Event i	Relation	Event j	Grammatical rules?	Cognitively logical?
1	A truck collided with a bus	cause	Four deaths	√	√
2	The scene is thick smoke rolling	accompany	Fire brigade sprinkles water to cool down the fire accident scene	√	√
3	A fire broke out on the 16th floor of Taisheng Building in Xu Chengguan District	follow	Investigation of the cause of the fire	√	√
4	According to local media reports in Indonesia	thought content	A strong earthquake with a magnitude of 6.3 occurred in the Pacific Ocean southwest of the country on the same day	√	√
5	A 5.8 magnitude earthquake hit the northern part of Chile on the 2nd of this month	cause	Shanxi's Fuping tremor felt significantly lasted 10 seconds	√	×
6	Aftershock occurred at the original epicenter	follow	Quickly mobilize the special duty Hong Kong North Hong Kong South Qiantang squadron of fire trucks and officers and soldiers	√	×

4.4. Event Causality Extraction Results with Different Models Using ERDAP

Causality is an important semantic relation among the multiple relations, which reflects the connection between events from cause to effect. Since texts contain more explicit or implicit causal knowledge, event causality extraction is important for tasks such as information retrieval, knowledge Q and A [2, 17], and event evolutionary graph construction [13], etc. However, due to the ambiguity and diversity of natural language texts, existing event causality extraction methods [4, 12, 24] transform the extraction problem into a classification problem by first performing feature extraction and then classification, which has the problem of insufficient semantic feature representation, proposed a method based on joint word vectors and Attention-Bidirectional Gate Recurrent Unit (Att-BiGRU) to address this problem. To verify the performance of the data augmentation method in this paper, we used four different models, Att-GRU, BiGRU, Att-BiLSTM, and Att-BiGRU, to perform event causality extraction on the CEC pre- and post-augmented datasets, and the results are compared in Table 2.

Table 2. Comparison of event causality extraction results on the pre- and post-augmented datasets (%).

Models	Precision	Recall	F1
Att-GRU	80.37	87.56	83.81
BiGRU	86.23	76.19	80.90
Att-BiLSTM	77.92	90.91	83.92
Att-BiGRU	92.25	85.19	**88.58**
Att-GRU $_{(+)}$	79.59	89.9	84.43
BiGRU $_{(+)}$	87.08	77.12	81.80
Att-BiLSTM $_{(+)}$	80.25	89.32	84.54
Att-BiGRU $_{(+)}$	92.97	87.18	**89.98**

4.5. Event Causality Extraction Results by the Same Model for Different Data Augmentation Methods

To fully measure the effectiveness of the ERDAP method proposed in this paper, four different data augmentation methods of synonym replacement (15%), random deletion (15%), contextual enhancement and ERDAP were used to generate event relation data in this section, and the samples before and after CEC data enhancement were extracted for event causality using the Att-BiGRU model with the best extraction performance mentioned above, and the results are shown in Table 3.

Table 3. Comparison of event causality extraction results before and after using data augmentation (%).

Methods	Precision	Recall	F1
Raw data	92.25	85.19	88.58
Synonym substitution (15%)	90.86	87.68	89.24
Random deletion (15%)	91.97	86.14	88.96
Contextual enhancement	92.22	86.97	89.52
ERDAP(ours)	92.97	87.18	**89.98**

According to the comparative analysis in Table 3, it can be seen that the F1-score of all four augmentation methods are improved, and compared with the improvement of about 0.5-1 point of the first three methods, the ERDAP method has the best effect with 1.4 points of improvement. Meanwhile, the random deletion method has the least improvement, indicating that deleting contextual words will expand the semantic space and thus improve the learning effect, but it also brings negative noise effects, while the synonym replacement method replaces words in the text using synonyms, which has some improvement but only considers word-level features without considering contextual context, and the contextual enhancement method uses a pre-trained language model that combines deeper semantic features, our approach incorporates deeper semantic features along with richer structural features.

5. Conclusions

Most existing methods that use event relational extraction as a supervised classification task suffer from insufficient labeled data. We propose a novel event relation data augmentation approach ERDAP that cleverly uses the relation prediction task for event relation data augmentation. Unlike previous approaches as back-translate texts and replace synonyms, we transform data augmentation into an event relation prediction problem. First, the event texts are converted into an event relational graph and the graph structure is constructed; then, the EventRGCN model is used to encode the relation graph, extract semantic features and structural features, and employ a semantic matching model for relation prediction; finally, a reasonable event relation triple is selected as the augmented data. The experimental results prove the effectiveness of the method, and the ERDAP method can improve event causality extraction by 1.4 percentage points on the CEC dataset. However, this paper only conducted experiments on the Chinese CEC dataset, and we will verify the performance of the method on more datasets and further investigate the event causality extraction method in the future.

References

[1] Balkus S. and Yan D., "Improving Short Text Classification with Augmented Data Using GPT-3," *Natural Language Engineering*, vol. 28, no. 6, pp. 1-30, 2022. doi:10.1017/S1351324923000438

[2] Cao Q., Trivedi H., Balasubramanian A., and Balasubramanian N., "DeFormer: Decomposing Pre-trained Transformers for Faster Question Answering," *arXiv Preprint*, arXiv: 2005.00697v1, pp. 1-11, 2020. https://doi.org/10.48550/arXiv.2005.00697

[3] Caselli T. and Vossen P., "The Event Story-Line Corpus: A New Benchmark for Causal and Temporal Relation Extraction," *in Proceedings of the Events and Stories in the News Workshop*, Vancouver, pp. 77-86, 2017. https://aclanthology.org/W17-2711

[4] Dasgupta T., Saha R., Dey L., and Naska 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, pp. 306-316, 2018. DOI:10.18653/v1/W18-5035

[5] Devllin J., Chang M., Lee K., and Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *in Proceedings of the NAACL-HLT*, Minneapolis, pp. 4171-4186, 2019. https://aclanthology.org/N19-1423.pdf

[6] Han S., Hao X., and Huang H., "An Event-Extraction Approach for Business Analysis from Online Chinese News," *Electronic Commerce Research and Applications*, vol. 28, no. 11, pp. 244-260, 2018. https://doi.org/10.1016/j.elerap.2018.02.006

[7] Hou Y., Liu Y., Che W., and Liu T., "Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding," *in Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, pp. 1234-1245, 2018. https://aclanthology.org/C18-1105

[8] Hussain Z., Gimenez F., Yi D., and Rubin D., "Differential Data Augmentation Techniques for Medical Imaging Classification Tasks," *Annual Symposium Proceedings Archive*, vol. 2017, pp. 979-984, 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977656/

[9] Kobayashi S., "Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations," in *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Conference*, New Orleans, pp. 452-457, 2018. DOI:10.18653/v1/N18-2072

[10] Krizhevsky A., Sutskever I., and Hinton G., "Imagenet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2012. https://doi.org/10.1145/3065386

[11] Li D., Mallinson J., Reddy S., and Lapata M., "Learning to Paraphrase for Question Answering," *in Proceedings of the Empirical Methods in Natural Language Processing Conference*, pp. 875-886, Copenhagen, 2017. DOI:10.18653/v1/D17-1091

[12] Li P. and Mao K., "Knowledge-Oriented Convolutional Neural Network for Causal Relation Extraction from Natural Language Texts," *Expert Systems with Application*, vol. 115, no. 3, pp. 512-523, 2019. https://doi.org/10.1016/j.eswa.2018.08.009

[13] Li Z., Zhao S., and Ding X., "EEG: Knowledge Base for Event Evolutionary Principles and Patterns," *in Proceedings of the 6th Chinese National Conference on Social Media Processing*, Beijing, pp. 40-52, 2017. https://doi.org/10.1007/978-981-10-6805-8_4

[14] Liu T., Cui Y., Yin Q., Zhang W., Wang S., and Hu G., "Generating and Exploiting Large-Scale Pseudo Training Data for Zero Pronoun Resolution," *in Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics, Vancouver, pp. 102-111, 2016. DOI:10.18653/v1/P17-1010

[15] Mirza P. and Tonelli S., "An Analysis of Causality between Events and its Relation to Temporal Information," *in Proceedings of the COLING the 25th International Conference on Computational Linguistics*, Dublin, pp. 2097-2106, 2014. https://aclanthology.org/C14-1198

[16] Schlichtkrull M., Kipf T., Bloem P., Berg R., Titov I., and Welling M., "Modeling Relational Data with Graph Convolutional Networks," *in Proceedings of the 15th International Conference on European Semantic Web*, Heraklion, pp. 593-607, 2018. https://doi.org/10.1007/978-3-319-93417-4_38

[17] Soares M. and Parreiras F., "A Literature Review on Question Answering Techniques, Paradigms and Systems," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 6, pp. 635-646, 2020. https://doi.org/10.1016/j.jksuci.2018.08.005

[18] Wei J. and Zou K., "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," *in Proceedings of the Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, pp. 6382-6388, 2019. https://aclanthology.org/D19-1670

[19] Wu X., Lv S., Zang L., Han J., and Hu S., "Conditional BERT Contextual Augmentation," *in Proceedings of the 19th International Conference on Computational Science*, Faro, pp. 84-95, 2019. https://doi.org/10.1007/978-3-030-22747-0_7

[20] Yang B., Yih W., He X., Gao J., and Deng L., "Embedding Entities and Relations for Learning and Inference in Knowledge Bases," *arXiv Preprint*, arXiv:1412.6575v2, pp. 1-12, 2015. https://www.semanticscholar.org/reader/8641230 6b777ee35aba71d4795b02915cb8a04c3

[21] Yu A., Dohan D., Luong M., Zhao R., Chen K., Norouzi M., and Le Q., "QANET: Combining Local Convolution with Global Self-Attention for Reading Comprehension," *in Proceedings of the 6th International Conference on Learning Representations*, Vancouve, pp. 1-16, 2018. https://openreview.net/pdf?id=B14TlG-RW

[22] Zavalsız M., Alhajj S., Sailunaz K., Ozyer T., and Alhajj R., "A Comparative Study of Different Pre-Trained DeepLearning Models and Custom CNN for Pancreatic Tumor Detection," *The International Arab Journal of Information Technology*, vol. 20, no. 3A, pp. 515-526, 2023. https://iajit.org/upload/files/doi-71684824197-24.pdf

[23] Zhang Y., Liu Z., Li Q., and Zhou W., "Construction of Event-Oriented Chinese Reference Corpus," *Journal of Shanghai University (Natural Science Edition)*, vol. 24, no. 6, pp. 900-911, 2018. doi:10.12066/j.issn.1007-2861.1888

[24] Zheng Q., Wu Z., and Zuo J., "Event Causality Extraction Based on Two-Layer CNN-BiGRU-CRF," *Computer Engineering*, vol. 47, no. 5, pp. 58-64, 2021. DOI:10.19678/j.issn.1000-3428.0057361

[25] Zuo X., Cao P., Chen Y., and Lui K., "Learn DA: Learnable Knowledge-Guided Data Augmentation for Event Causality Identification," *arXiv Preprint*, arXiv:2106.01649v1, pp. 1-14, 2021. file:///C:/Users/user/Downloads/LearnDA_Learnable_Knowledge-Guided_Data_Augmentati.pdf

Ruijuan Hu she was born in 1983. She is a Bachelor of Engineering from Jilin Institute of Architecture and Engineering, and a Master of Engineering from Changchun University of Science and Technology. She works as an associate professor at PLA Strategic Support Force Information Engineering University, with 14 years of teaching experience. She is currently pursuing a PhD in software engineering at the university. Her research interests are Natural Language Processing and Knowledge Engineering.

Haiyan Liu he was born in 1971. He graduated from PLA Strategic Support Force Information Engineering University with a PhD. He is a professor and have guided dozens of graduate and doctoral students. His research interests are Spatio-Temporal Data Mining and Big Data Integration and Governance.

Wentao Yu he was born in 1988. He is a Bachelor and a Master of Engineering from PLA Strategic Support Force Information Engineering University. He got a PhD in 2021 and now works at PLA Strategic Support Force Information Engineering University. He is also a postdoctoral fellow at the Chinese Academy of Science from 2023 and his research interests are Natural Language Processing and Event Extraction.