An Interactive Attention-Based Approach to Document-Level Relationship Extraction

Zhang Mei, Zhao Zhongyuan, Xu Zhitong
School of Information Science and Technology, North China University of Technology, Beijing 100144, China

***Abstract*—Document-level relation extraction entails sifting through extensive document data to pinpoint relationships and pertinent event details among various entities. This process aids intelligence analysts in swiftly grasping the essence of the content while revealing potential connections and emerging trends, thus proving invaluable for research purposes. This paper puts forward a method for document-level relation extraction that leverages an interaction attention mechanism. Initially, building on an evidence-based approach for extracting relations at the document level, the interaction attention mechanism is introduced, extracting the final layer of hidden states containing rich semantic information from the document encoder. Subsequently, these concealed states are fed into a self-attention layer informed by dependency parsing. The outputs from both elements serve as distinct supervisory signals for the interactive input. By pooling these output results, it can derive context embeddings that possess enhanced representational power. Preliminarily, relation triples are extracted using the relation classifier. In conclusion, building on the preliminary relationship results, the process of relationship inference is carried out independently using pseudo-documents created from the source material and pertinent evidence. Only those relationships with a cumulative inference score that surpasses a certain threshold are regarded as the final outcomes. Experimental findings from the publicly accessible datasets indicate commendable performance.**

Keywords—Document-level relation extraction; interaction attention-based; the baseline model

I. Introduction

Conventional relation extraction models typically focus on individual sentences, overlooking the subtle contextual and semantic connections that exist between sentences throughout the entire document [1]. Document-level relation extraction is about finding and understanding the connections between different parts of a document. For instance, in document-level relation extraction, a relationship between entities might span multiple sentences, making it challenging for the model to accurately extract those relationships. Fig. 1 from the DocRED dataset shows part of a document. While the relationship between "The Legend of Zelda" and "Capcom and Flagship" may seem clear, the text has many connections that require analysis. The Legend of Zelda was created by Capcom and Flagship with guidance from Nintendo. But a more thorough examination shows that Nintendo plays a big role. They guide the design choices and overall development of the game. This scenario necessitates that the model possess robust competencies in discerning underlying connections and executing reasoning in an efficacious manner [2].

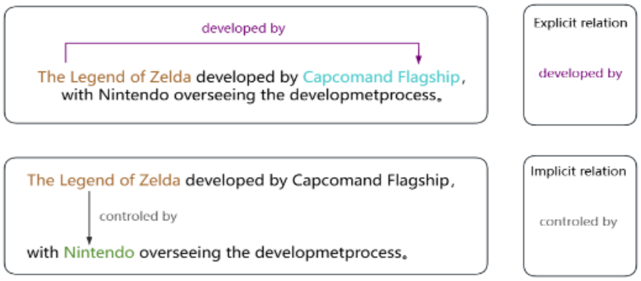
developed by
The Legend of Zelda developed by Capcomand Flagship,
with Nintendo overseeing the developmetprocess。
Explicit relation
developed by
The Legend of Zelda developed by Capcomand Flagship,
controled by
with Nintendo overseeing the developmetprocess。
Implicit relation
controled by

Fig. 1. Sample extraction of displayed and implicit relations under the DocRE task.

The proposed model must possess the robust capability to comprehend contextual information meticulously, refraining from concentrating on isolated sentences; it should analytically encompass the comprehensive document. The sophisticated model articulated uniquely embodies essential evidence data within the enriched supplemental information stream, utilizing this crucial input to augment relational reasoning seamlessly, thus significantly strengthening the holistic understanding of extensive documents by the model. Henceforth, numerous inadequacies remain when tackling profoundly intricate implicit connections comprehensively. Therefore, it is proposed to use interaction attention to enable the model to fully grasp contextual information, thereby further enhancing the model's reading comprehension ability for the entire document [3].

An entity in a document may be mentioned many times, but not always with the full name. It may be in the form of an abbreviation, acronym, or code name. Abbreviations and code names are analyzed on the DocRED, Re-DocRED, and CDR datasets. For example, 61.1% of the relationship instances in the same document in the DocRED dataset need to be recognized for reasoning, and only 38.9% of the relationship instances can be extracted by simple pattern recognition. This shows that commonly used pre-trained language models (e.g., Transformer, Bert, etc.) cannot completely solve the long-distance dependency and improve the overall understanding of the implicit structure of documents.

Therefore, using interaction attention to find more hidden information in the document can help document-level relation extraction models. Based on evidence-driven document-level relation extraction methods, and to mine deeper information in the document to enhance relation reasoning [4], this paper proposes using interaction attention mechanisms during document modeling to help the model uncover hidden information in the document.

II. Model Construction

To address the issues, a new document-level relation extraction model is proposed. This approach uses BERT, a pre-trained language model, with interaction attention to improve how it extracts information from data after completing evidence sentences. It also lets each attention head learn features from different parts of the document, which extends the model's capabilities. This approach helps the model learn about the whole context and makes up for the fact that attention mechanisms can only learn about what is right in front of them.

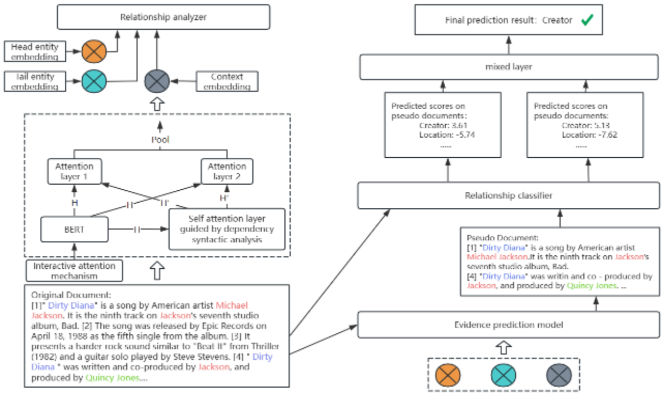
Relationship analyzer
Head entity embedding
Tail entity embedding
Context embedding
Pool
Attention layer 1
Attention layer 2
H
H'
H
H'
BERT
H
Self attention layer guided by dependency syntactic analysis
Interactive attention mechanism
Original Document:
[1]" Dirty Diana" is a song by American artist Michael Jackson. It is the ninth track on Jackson's seventh studio album, Bad. [2] The song was released by Epic Records on April 18, 1988 as the fifth single from the album. [3] It presents a harder rock sound similar to "Beat It" from Thriller (1982) and a guitar solo played by Steve Stevens. [4] " Dirty Diana " was written and co-produced by Jackson, and produced by Quincy Jones....
Final prediction result: Creator ✓
mixed layer
Predicted scores on pseudo documents: Creator: 3.61 Location: -5.74 ......
Predicted scores on pseudo documents: Creator: 5.13 Location: -7.62 ......
Relationship classifier
Pseudo Document:
[1] "Dirty Diana" is a song by American artist Michael Jackson.It is the ninth track on Jackson's seventh studio album, Bad.
[4] "Dirty Diana" was writin and co - produced by Jackson, and produced by Quincy Jones. ...
Evidence prediction model

Fig. 2. IADocRE model diagram.

Fig. 2 shows how the model uses interaction attention to find information in documents. The model has two parts: relation extraction (on the left) and relation extraction and reasoning (on the right).

A. Document Level Encoder

A document, designated as d, is constituted of N sentences $\{S_n\}$, L Token tokens $\{h_l\}$, E named entities $\{e_i\}$ and all proper noun mentions $\{m_{ij}\}$ for each entity. The objective of document-level relationship extraction is to ascertain the set of potential relationships between all pairs of entities from a predefined set of relationships R, which is provided as $R \cup \{NR\}$, where NR denotes the absence of a relationship[5]. It should be noted that an entity may be referenced on more than one occasion within a single document. Consequently, for each entity e_i, there can be multiple mentions $\{ m_j^i \}$ Ne_{ij}=1. In the absence of a relationship between the entities in the pair (e_h, e_t), the designation NR is applied.

And make the head entity and the tail entity eh and et respectively during the test period, all entity pairs $(e_h, e_t)_{h, t \in (1...n), h \neq t}$. Essentially this is a multi-labeling classification problem because there may be multiple relationships between e_h and e_t. If the relation r exists between (e_h, e_t), it is initially classified as a positive class PT, otherwise it belongs to the negative class NT. For each entity pair of the NR relation (e_h, e_t), its evidence V_{ht} is defined as the subset of sentences in the document from which the relation can be inferred.

The document is initially encoded using BERT, which captures contextual information and semantic representations. Compared to the sentence-level encoder value focusing on the encoding of individual sentences, the document-level encoder has an inherent advantage in dealing with the task of document-level relation extraction [6]. Specifically, for a given document $d=[h_l]$, special tags are added before and after the mention of each entity using the mainstream method, i.e., [CLS] + ‘Entity’ + [SEP], and then encoded using an encoder:

$$H, A = \mathrm{PrLM}[h_1,...,h_L] \tag{1}$$

where, the hidden layer state of the last layer is usually denoted as H^L, and L denotes the number of Lth layer in the BERT model.

B. Dependent Syntax Guided Self-Attention Layer

Xu et al. found that when a language model like BERT learns text, it ignores words not in the subject, predicate, or object. This research suggests that it is crucial to try to improve pre-trained language models by giving some of the attention to words other than subject, predicate, and object. The technique of dependency parsing is used to guide the model's attention towards other information in the sentence. The goal of dependency parsing is to help BERT focus on information other than the subject and predicate. In addition to using the self-attention layer to enhance attention, the incorporation of an interactive attention mechanism ensures that the enhanced model is capable of recognizing more contextual information than the original BERT model.

The text is first contextualized using the language model BERT [7]. Google found that the 12-layer Transformer encoder architecture is the best for classification. The BERT encoder captures contextual information in the text through a multi-head self-attention mechanism. The Feedforward Neural Network (FNN) performs non-linear transformations and is combined with the BERT encoder through Residual Connection and Normalization. The BERT encoder captures context in text through a self-attention mechanism. The FNN performs nonlinear transformations, which are combined and merged by residual connection and normalization. The BERT encoder was designed to improve classification effectiveness, storing rich contextual information in the document. It makes no difference if the last two or three layers are used for comparison. The last hidden layer state H is used as the initial representation of the sentence. The initial representation H of the sentence is passed to the self-attention layer, guided by the dependent syntactic analysis. The output is denoted as H'. H' shows which weights in the initial representation H are kept by the self-attentive layer. This is done according to the structure of the dependent syntactic analysis tree. Only the weights that contain dependency relations are kept. The weights of inflectional words, dummy words, and quantifiers that are not related are reduced. These can mislead the model or distract the attention weights.

Dependency parsing is about finding the links between words in a document. In Section II (B), dependency is defined as a relationship between words. Every sentence has a main word that is connected to other words.

Specifically, set the length of a sentence Sen in the document to be n, and the set of words in the sentence to be $\{W\} = \{w_1, w_2, w_n\}$, determine the ancestor set {P} of all nodes through a dependency analysis tree and create an n × n dimensional MASK matrix, denoted as M[8]. The set P consists of the directly or indirectly dominating subsets of any word w_i.

Specifically, for any word w_i ($i \in$ Sen) within a sentence, if a word w_j ($j \in$ Sen) appears in the ancestor set {P}, then the value of row i and column j of the MASK matrix M is set to 1, and all the other column positions within row i are set to 0. The computation formula is as follows:

$$\mathrm{M[i, j]} = \begin{cases} 1 & if \quad j \in P_i \quad or \quad j = i \\ 0 & otherwise \end{cases} \tag{2}$$

In this study, the attention mechanism has three parts: the query vector Query (Q), the key vector Key (K), and the value vector Value (V), where K is usually set equal to V. The query vector and the key vector determine the weighting coefficients of the value vector. The query vector Q and the key vector K determine the weighting coefficients of V. Self-attention has the same query vector Q as the key vector K. Multi-head attention splits the query vector Q into parts and extracts multiple K-V pairs from the text. In conclusion, the above features are combined.

The final hidden layer state H of BERT is input into the multi-head self-attention layer, and the MASK matrix M of the previous complementary information is used to dot-multiply the query vector Q and the key vector K. By calculating the weights in this way, the final attention representation W_i, with i denoting each attention head, can be obtained. The computational formula is as follows:

$$\mathrm{A}_{\mathrm{i}}^{'} = soft\max(\frac{M(Q_i^{'} K_i^{'})}{\sqrt{d}}) \tag{3}$$

$$\mathrm{W}_{\mathrm{i}}^{'} = \mathrm{A}_{\mathrm{i}}^{'} \mathrm{V}_{i}^{'} \tag{4}$$

The output W_i from each attention head is merged and the result is output to the FNN network, where it is activated with the GeLU. It is then passed again to another FNN network. After ths series of operations, the representation H' can be obtained following guidance by the dependency parsing after passing through the Normalization process.

C. Interactive Attention Layer

The initial representation H and the representation H' are obtained after being guided by dependent syntactic analysis. The two outputs are then fused to enhance understanding between the two representations by the model. As shown in the figure, the initial representation H of the sentence is used as the key vector Key and value vector Value of the attention layer 1, and the representation H' guided by dependent syntactic analysis is used as the query vector Query of the attention layer 1, both of them are subjected to softmax operation, and the output result is the representation vector W_1 of the attention layer 1; similarly, the output result of the attention layer 2 is the representation W_2.The computation of the two attention layers is carried out at the same time, and the computation formula is as follows:

$$\mathrm{W}_1 = \mathrm{softmax}(\frac{Q_1 K_1^{\mathrm{T}}}{\sqrt{d_{k_1}}}) \mathrm{V}_1 \tag{5}$$

$$\mathrm{W}_2 = \mathrm{softmax}(\frac{Q_2 K_2^{\mathrm{T}}}{\sqrt{d_{k_2}}}) \mathrm{V}_2 \tag{6}$$

Fusing W_1 with W_2 to obtain the output result of the interactive attention layer representation W_3, representation W_3 enables the overall model to pay more attention to the semantic information in the sequence related to the current position in the output. Equation is shown in 4- 7, α is used to balance the parameters of W_1 and W_2, where the value of α is 0.5.

$$\mathrm{W}_3 = \alpha \mathrm{W}_1 + (1 - \alpha) W_2 \tag{7}$$

The BERT pre-trained language model prescribes, before input, the need to mark specific entities or sentences with special symbols in front of them and the endings with special symbols[9], i.e., [CLS] and [SEP]. The essence of [CLS] denotes the synthesized information of the whole sentence, so only the initially vector h0 of the W_3 vector is taken out and the attention A' of the whole sentence is computed using the soft-max function:

$$\mathrm{A}^{'} = \mathrm{softmax}(\mathrm{W}_3[0]) \tag{8}$$

Before the document is entered into BERT, each entity e_i is required to use the embedding of the special symbol as its mention embedding $\mathrm{m}_{\mathrm{j}}^{\mathrm{i}}$. Subsequently, the embedding of entity e_i over all its mention embeddings is obtained by employing Log-Sum-Exp pooling:

$$\mathrm{e}_{\mathrm{i}} = \log \sum_{j} \exp(m_j^i) \tag{9}$$

To predict the relationships between different entity pairs, the model may need to focus on different parts of the context. To capture the contextual dependencies associated with each entity pair (e_h, e_t), its contextual embedding is computed based on the interaction attention A':

$$\mathrm{C}_{\mathrm{h,t}} = H^{T} \frac{\mathrm{A}_{\mathrm{h}}^{'} \circ \mathrm{A}_{t}^{'}}{(\mathrm{A}_{\mathrm{h}}^{'})^{\mathrm{T}} \mathrm{A}_{t}^{'}} \tag{10}$$

where ^{0} denotes the Hadamard product, and A_h is the head entity's attention to all the tokens in the document, obtained by leveling out the mentions of the head entity. A_t is the same. Tokens that are highly attentive to both e_h and e_t must necessarily be important to both head and tail entities, and so should have more interactions on context embedding.

D. Classification of Relationships

To predict the relationship between entity pairs (e_h, e_t), the model initially computes their context-aware representations (z_h, z_t) by combining their entity embeddings (e_h, e_t) with their context embeddings $c_{h,t}$, and then utilizes a bilinear function to compute the logit of the likelihood of the existence of a relationship $r \in R$ between e_h and e_t.

$$\mathrm{Z}_{\mathrm{h}} = \tanh(W_h e_h + W_{c_h} C_{h,t}) \tag{11}$$

$$\mathrm{Z}_{\mathrm{t}} = \tanh(W_t e_t + W_{c_t} C_{h,t}) \tag{12}$$

$$\mathrm{y}_{\mathrm{r}} = \mathrm{Z}_{\mathrm{h}} \mathrm{W}_{\mathrm{r}} \mathrm{Z}_{\mathrm{t}} + \mathrm{b}_{\mathrm{r}} \tag{13}$$

where W_h, W_t, W_{ch}, W_{ct}, W_r, b_r are learnable parameters. Since the model may have different confidence levels for different entity pairs, an adaptive threshold loss is used, which learns a virtual relationship class TH that serves as a dynamic threshold for each entity pair:

$$y_{TH} = z_h W_{TH} z_t + b_r \tag{14}$$

In the inference process, for each tuple (eh, et, r), the predicted score is obtained:

$$S^{(O)}_{h,r,t} = y_r - y_{TH} \tag{15}$$

In conclusion, the training objective for relation extraction is defined as follows:

$$L_{RE} = -\sum_{r\in P^T_{h,t}} \log(\frac{\exp(y_r)}{\sum_{r'\in P^T_{h,t}\cup\{TH\}} \exp(y_{r'})}) - \log(\frac{\exp(y_{TH})}{\sum_{r'\in N^T_{h,t}\cup\{TH\}} \exp(y_{r'})}) \tag{16}$$

Where $P^T_{h,t}$ indicates that a relationship exists between two entities and $N^T_{h,t}$ indicates that no relationship exists between entities.

E. Relational Reasoning

If the evidence sentences contain all the relevant information, there's no need to use the entire document for relationship extraction. No system can extract 100% of the evidence without omitting some sentences. Relying on extracted evidence alone may miss important information in the document, which could affect performance. The original document and the extracted evidence are combined to get the prediction results. If there are no evidence annotations, the results can be learned by the evidence prediction model from Section III or extracted by the auxiliary experiments in Section III.

Specifically, as shown in Fig. 1, a set of relationship prediction scores are initially obtained from the original document $S^{(O)}_{h,r,t}$. Then a pseudo-document d' is constructed for each entity pair by concatenating the extracted evidence sentences V_{ht} in the order in which they appear in the original document. The prediction score of the relational extraction model for the pseudo-document is recorded as $S^{(P)}_{h,r,t}$. In conclusion, fusion results are obtained by aggregating the two sets of predictions through a hybrid layer:

$$P_{Fuse(r|e_h,e_t)} = (S^{(O)}_{h,r,t} + S^{(P)}_{h,r,t} - \tau) \tag{17}$$

τ denotes the balancing parameter of the source document and the pseudo-document. The final loss function is as follows:

$$L_{Fuse} = -\sum_{r\in R} y_r \cdot P_{Fuse(r|e_h,e_t)} + (1-y_r)\cdot \log(1-P_{Fuse(r|e_h,e_t)}) \tag{18}$$

If $y_r = 1$ indicates that a relationship exists between the entities. Conversely, $y_r = 0$ indicates that no relationship exists between the entities.

III. Experimental Results and Analysis

The commonly used datasets for document-level relation extraction, such as DocRED, Re-DocRED, CDR, and GDA, are employed to investigate the performance differences between the model with the introduction of the interaction attention mechanism and the baseline model.

A. Data Sets

The experiments are mainly evaluated on DocRED, CDR, GDA and Re-DocRED datasets. Since CDR and GDA are relationally extracted datasets under the medical domain, since both datasets do not have annotation information such as evidence sentences, CDR and GDA are placed in the same representation for comparison [11]. Re-DocRED and DocRED tend to be generalized domain relationally extracted datasets, both containing evidence sentence information and remote supervision data, therefore CDR and GDA datasets are set in one group for comparison and DocRED and Re-DocRED are set in one group for comparison. The statistical information of the four datasets is shown in Table I.

TABLE I. Data Set Information

Statistical information	CDR	GDA	DocRED	Re-DocRED
Training Documentation	500	23353	3053	3053
verification document	500	5839	1000	1000
test document	500	1000	1000	1000
predefined relationship	2	2	97	97
Average number of entities	7.6	5.4	19.5	19.6
Average number of sentences	9.7	10.2	8.0	8.1

B. Evaluation Indicators and Parameterization

1) Evaluation indicators: As in the previous section, Precision (P), Recall (R) and F1 values were used as evaluation metrics for the experiment, which were calculated as follows:

$$P = \frac{TP}{TP+FP} \tag{19}$$

$$R = \frac{TP}{TP+FN} \tag{20}$$

$$F_1 = \frac{2\times P\times R}{P+R} \tag{21}$$

In addition to the above commonly used evaluation metrics, there also exists the I_{gn} F_1 evaluation metric in document-level relational extraction. I_{gn} F_1 was proposed by Yao et al. The model learns relational facts existing in the training set during the training phase, which is shared with the validation set and the test set. Then if the model has learned certain relationship facts in the training set, it will inevitably affect the model's judgment in the validation or testing phase [12]. This approach obviously produces immeasurable interference in model performance evaluation, so in order to have a fairer evaluation index, after removing the shared relationship facts in the training and validation sets and the test set, the performance evaluation of the document-level relationship extraction model is re-conducted.

2) *Parameter setting:* The IADocRE model is based on Pytorch and Huggfacing's Transformer implementation, and the model is encoded using bert as a pre-trained language model [9]. The experiments used AdamW as the optimizer in the DocRED experiments. During the self-training phase of evidence sentences, the learning rate was set to 5e-5, Warmup was set to 0.06, and the Dropout rate was set to 0.1. When running the whole model, the experiments were trained and evaluated on four RTX 3090 24GB GPUs. The specific experimental parameters are shown in Table II.

C. Main Experiment and Analysis of Results

1) Experimental results and analysis of DocRED and Re-DocRED datasets: The main results of the IADocRE model for the two datasets are shown in Table III, the IADocRE model achieved 65.54 and 63.76 on F_1 and I_{gn} F_1 of the DocRED dataset, and 79.43 and 79.05 on F_1 and I_{gn} F_1 of the Re-DocRED dataset, and the scores on the two datasets have exceeded those of the existing baseline models. The experimental results prove that the document-level relationship extraction method based on interactive attention proposed in this paper is effective. From the table of experimental results, it can be observed that among the two mainstream methods for Document-level Relation Extraction, the effect of the model based on the sequence method is usually superior to that based on the graph method.

Table III and Table IV shows that the IADocRE model is ahead of the existing baseline model level. The I_{gn} F_1 evaluation metrics are missing for some of the models in the experimental results because some of the models were not experimentally validated on the latest dataset at the time of publication, and if the model is reproduced and validated on the latest dataset, the experimental process is again affected by the initial parameter settings, the type and number of hardware devices, and the choice of optimizer strategy. Therefore, partial results on the Re-DocRED dataset are denoted by '-'.

TABLE II. OPTIMAL PARAMETER SETTINGS FOR THE EXPERIMENT

point	**parameters**	**(be) worth**
Train	Warmup	0.06
	lr	5e-5
	evi_thresh	0.2
	Dropout	0.1
	max_grad_norm	1.0
Fine-tune	Warmup	0.06
	lr	1e-6
	evi_thresh	0.2
	Dropout	0.1
	max_grad_norm	2.0

TABLE III. EXPERIMENTAL RESULTS OF IADOCRE MODEL ON DOCRED AND RE-DOCRED DATASETS (%)

Model		**DocRED**				**Re-DocRED**			
		Dev		***Test***		***Dev***		***Test***	
		F1	***Ign F1***	***F1***	***Ign F1***	***F1***	***Ign F1***	***F1***	***Ign F1***
Graph-based Models	AGGCN	52.47	46.29	51.45	48.89	-	-	-	-
	LSR-BERT	59.00	52.43	59.05	56.97	-	-	-	-
	GLRE-BERT	-	55.40	-	57.40	-	-	-	-
	GCGCN-BERT	57.35	55.43	56.67	54.53	-	-	-	-
	GRACR-BERT	59.73	57.85	58.54	56.47	-	-	-	-
	HeterGSAN	60.18	58.13	59.45	57.12	-	-	-	-
Transformer-based Models	BERT	54.16	-	53.20	-	-	-	-	-
	BERT-Two-Step	54.42	-	53.92	-	-	-	-	-
	HIN-BERT	56.31	54.29	55.60	53.70	-	-	-	-
	CoreBERT	57.51	55.32	56.96	54.54	-	-	-	-
	SSAN-BERT	59.19	57.03	58.16	55.84	-	-	-	-
	RSMAN-BERT	59.25	57.22	59.29	57.02	-	-	-	-
	JEREX	62.24	60.39	62.15	60.29	74.77	73.34	74.79	73.48
	ATLOP-BERT	61.01	59.11	61.30	59.31	79.29	78.32	79.46	78.52
	EIDER (Rule)-BERT	62.34	60.36	62.21	60.23	-	-	-	-
	EIDER-BERT	62.48	60.51	62.47	60.42	-	-	-	-
	DocuNET	65.25	63.22	65.26	**63.23**	78.90	78.20	78.99	78.28
	IADocRE-BERT (ours)	**65.54**	**63.76**	**65.27**	63.16	**79.43**	**79.05**	**79.38**	**79.21**

In addition to the statistics on the ternary group prediction results, the experiments also included statistics on the evidence prediction results. According to the work of Huang [6] and Xie [10] , it is known that only E2GRE and EIDER have published methods for evidence extraction results. As shown in Table IV, IADocRE's method significantly outperforms E2GRE and EIDER on BERT, the evidence sentence extraction results are improved by 5.13 and 1.56 on the validation set, and by 4.51 and 1.59 on the test set, respectively in terms of the score metrics, but, there is still a lot of room for improvement in the evidence extraction.

TABLE IV. RESULTS OF EVIDENCE EXTRACTION EXPERIMENTS OF THE IADOCRE MODEL ON THE DOCRED DATASET (%)

Model	**Evi F1**	
	Dev	***Test***
E2GRE-BERT	47.14	48.35
EIDER-BERT	50.71	51.27
IADocRE-BERT (ours)	**52.27**	**52.86**

2) Experimental results and analysis of CDR and GDA datasets: The main results of the IADocRE model for the two datasets are shown in Table V. There is no concept of shared relational facts in the training and validation sets of the CDR and GDA datasets, so the evaluation metrics for both datasets are only F1.The IADocRE model achieves an F1 of 78.2 for the CDR dataset, and 87.8 for the GDA dataset. The performance on the GDA dataset has already exceeded that of existing benchmark models, but there remains a gap in performance on the CDR dataset, with the score being close to that of the SAIS model. Comparison of the two models reveals that: the SAIS model focuses more on the intermediate step of supervising and enhancing the model, which is a method that can more accurately capture the relevant context and entity-type information for the combination of the entity-type information and the evidence for achieving the effect of data enhancement. The intermediate steps of the supervised training process, such as Coreference Resolution, Entity Recognition and evidence-based retrieval, are clarified to help the model learn better.

TABLE V. EXPERIMENTAL RESULTS OF IADOCRE MODEL ON CDR AND GDA DATASETS (%)

Model	**CDR**	**GDA**
BERT	65.1	82.5
LSR-BERT	65.9	82.2
DHG-BERT	65.9	83.1
SSAN-BERT	68.7	83.7
GLRE-BERT	68.5	-
ATLOP-BERT	69.4	83.9
SIRE-BERT	70.8	84.7
DocuNET-BERT	76.3	85.3
SAIS-BERT	**79.0**	87.1
IADocRE-BERT (ours)	78.2	**87.8**

However, the SAIS model has the issues of increased complexity, higher data and annotation requirements, and greater consumption of computational resources compared to the IADocRE model. In summary, IADocRE is able to extract more ternary information under the medical domain dataset.

D. Ablation Experiments

1) Experimental results and analysis of ablation of DocRED and Re-DocRED datasets:

a) Analysis of the impact of interactive attention mechanisms on model performance:

All other things being equal, the contribution of the interactive attention mechanism to the model is explored by conducting multiple experiments on pre-trained language models with and without the introduction of interactive attention, respectively. The results of the ablation experiments on the DocRED and Re-DocRED datasets are shown in Tables IV-VI. When the overall model removes the interactive attention mechanism and only uses the original BERT, there is a significant decrease in the extraction effect, with F_1 and I_{gn} F_1 decreasing by 1.73 and 1.86 on DocRED and the results on Re-DocRED decreasing by 1.79 and 1.49, respectively. The ablation experiments performed on the interactive attention show that the overall model metrics decrease the most, and it can be inferred that the interactive attention mechanism is very effective in enhancing the model extraction performance.

To explore the effect of the dependent syntactic bootstrap attention layer on the model in more detail, the dependent syntactic analysis was replaced with an ordinary self-attention layer [13]. The model's extraction performance showed a decrease of 0.77 in the F_1 metric and 1.57 in I_{gn} F_1 on the DocRED dataset, indicating that the IADocRE model relies heavily on the dependent syntactic analysis attention layer, especially when ignoring relational facts. The performance significantly drops when replaced with a standard attention layer.The F_1 metric on the Re-DocRED dataset decreases by 1.04 and the I_{gn} F_1 decreased by 0.89. Considering Tan et al.'s revision of the Re-DocRED dataset, which removed a large amount of shared relational facts, the observed decreases in metrics are within the normal range. In summary, the experiments demonstrate the effectiveness of using the interactive attention mechanism in the document-level relationship extraction task.

b) Analysis of the impact of source documents and pseudo-documents on model performance:

After the ablation experiments with the interactive attention mechanism, ablation experiments were also conducted on the pseudo-document and source document parts. To explore whether the presence of source documents and pseudo documents in the relational reasoning part helps the model improve the effectiveness of relational reasoning [14]. As shown in Table IV and VI, when the model retains the interactive attention mechanism and removes the pseudo-document and only retains the source document part for relational reasoning, there is a significant decrease in the effect. When the model retains the interactive attention mechanism and removes the source document for relational reasoning, the model's F_1 and I_{gn} F_1 metrics in DocRED are 0.65 and 0.7, respectively. Observing the experimental data of the source document and pseudo-documents' ablation in the table. It can be seen that, by removing the pseudo-document, the model's F_1 and I_{gn} F_1 metrics in DocRED decrease by 1.12 and 1.32, respectively. The ablation experimental results for both the source and the pseudo-document show that: source document and pseudo-document have a significant decrease in their effectiveness. The results of the ablation experiments on source documents and pseudo-documents show that the model is more inclined to reason on pseudo-documents than on source documents, but it cannot completely rely on pseudo-documents for reasoning. Alternatively, the pseudo-document occupies a higher position in the model's reasoning.

2) Results and analysis of ablation experiments on CDR and GDA datasets:

a) Analysis of the impact of interactive attention mechanisms on model performance:

The experimental setup is the same as in the previous subsection, and the ablation experiments are conducted using the pre-trained language model with and without the introduction of the interactive attention mechanism, respectively. As shown in Table VII, using only BERT as the pre-trained language model decreases the extraction effect by 1.6 on the CDR dataset and 1.8 on the GDA dataset. The phenomenon suggests that the introduction of the interactive attention mechanism is effective in directing the model to focus on implicit expressions in biomedical documents, such as the roles of chemicals and diseases or the associations of genes with diseases, which are not always directly explicitly mentioned. Therefore, the interactive attention mechanism is effective in deepening the understanding of biomedical domain knowledge and context.

In order to explore the effect of the dependent syntactic analysis-guided attention layer on the overall model, this layer was replaced with the ordinary attention layer for experimental analysis. As shown in Tables IV- VII, the extraction effect on CDR and GDA decreased by 0.8 and 0.6, respectively, and this result indicates that the dependent syntactic analysis-guided attention layer helps the model to improve its comprehension of logical and causal relationships between sentences. In summary, the interactive attention mechanism is effective for the model's deep understanding of complex terminology and descriptive processes in biomedicine.

b) Analysis of the impact of source documents and pseudo-documents on model performance:

As can be seen in Tables IV- VII, retaining the source document for ablation experiments results in a decrease in CDR and GDA of 1.2 and 1.1, respectively. Retaining the pseudo-document for ablation experiments results in a decrease in CDR and GDA of 1.4 and 1.5, respectively. This suggests that the model reasoning partially relies more on the source document. The reason for this is that relational extraction in the biomedical domain often requires inference at multiple levels, e.g., understanding how a specific chemical can cause a change in disease state through a specific biological pathway may require inference across multiple parts of the document. If only pseudo-documents are retained for relational reasoning, there is a high risk of losing important information at other levels. Therefore, in biomedical domain datasets, the extraction performance is better when the model is capable of in-depth understanding and reasoning about the complex interactions between entities [15].

TABLE VI. RESULTS OF ABLATION EXPERIMENTS OF IADOCRE MODEL ON DOCRED AND RE-DOCRED DATASETS

Model		DocRED		Re-DocRED	
		F1	***Ign F1***	***F1***	***Ign F1***
$\text{IADocRE}_{\text{all}}$		**65.54**	**63.76**	**79.43**	**79.05**
$\text{IADocRE}_{\text{w/o-IA}}$		63.81	61.90	77.64	77.56
$\text{IADocRE}_{\text{w/o-DP}}$		64.77	62.19	78.39	78.16
Document	-NoPseudo	64.42	62.44	78.87	78.41
	-NoOrigdo	64.89	63.06	79.02	78.92
	-all	64.40	62.38	78.94	78.50

TABLE VII. RESULTS OF ABLATION EXPERIMENTS OF THE IADOCRE MODEL ON CDR AND GDA DATASETS

Model		CDR	GDA
$\text{IADocRE}_{\text{all}}$		**78.2**	**87.8**
$\text{IADocRE}_{\text{w/o-IA}}$		76.6	86.0
$\text{IADocRE}_{\text{w/o-DP}}$		77.4	87.2
Document	-NoPseudo	77.0	86.7
	-NoOrigdo	76.8	86.3
	-all	76.6	86.2

IV. CONCLUSION

When the document-level relational extraction model fills in missing evidence information, the performance shows improvement compared to the baseline model, but there is still much room for enhancement. For example, the presence of implicit information in documents requires the model to have strong reasoning capabilities. To address this problem, the introduction of the interactive attention mechanism can help the model understand the semantic and contextual information of documents from a global perspective. By combining document information and evidence information, the model can supplement entity relationship information, thus improving the accuracy and completeness of relationship extraction. The introduction of the interactive attention mechanism helps the model to better understand the deep logical relationships, which improves the performance of the model in relational reasoning.

ACKNOWLEDGMENT

This work was supported by the Humanities and Social Science Foundation of the Ministry of Education (Grant No.21YJA740052) and National Natural Science Foundation of China (62476007).

REFERENCES

[1] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph LSTMs. Transactions of the Association for Computational Linguistics, 5:101-115.

[2] Patrick Verga, Emma Strubell, and Andrew McCallum.2018. Simultaneously self-attending to all mentions for full-abstract biological

relation extraction. in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, V olume 1 (Long Papers), pages 872-884, New Orleans, Louisiana. association for Computational Linguistics.

[3] Ma Y, Wang A, Okazaki N. DREEAM: Guiding Attention with Evidence for Improving Document-Level Relation Extraction[J]. arXiv preprint arXiv:2302.08675, 2023.

[4] Kevin Huang, Peng Qi, Guangtao Wang, Tengyu Ma, and Jing Huang. 2021a. Entity and evidence guided document-level relation extraction. in Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pages 307-315, Online. .

[5] Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In Proceedings of the AAAI Conference on Artificial Intelligence.

[6] Yiqing Xie, Jiaming Shen, Sha Li, Y uning Mao, and Jiawei Han. 2022. eider: empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In Findings of the Association for Computational Linguistics: ACL 2022, pages 257-268, Dublin, Ireland. Association for Computational Linguistics.

[7] Benfeng Xu, Quan Wang, Y ajuan Lyu, Y ong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: modeling mention dependencies for document -In Proceedings of AAAI.

[8] Robin Jia, Cliff Wong, and Hoifung Poon. 2019. document-level n-ary relation extraction with multi-scale representation learning. in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. in Proceedings of NAACL.

[10] Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. double graph based reasoning for document-level relation extraction. in Proceedings of EMNLP.

[11] Ding Xiao. Research on document-level relationship extraction technique[D]. Strategic Support Forces Information Engineering University,2023.DOI:10.27188/d.cnki.gzjxu.2023.000018.

[12] Zhu Taojie. Research on Document-Level Relationship Extraction Techniques for Long-Range Entities[D]. Strategic Support Forces Information Engineering University,2023.DOI:10.27188/d.cnki.gzjxu.2023.000079.

[13] Xiao C, Yao Y, Xie R, et al. Denoising relation extraction from document-level distant supervision[J]. arXiv preprint arXiv:2011.03888, 2020.

[14] Zhu Taojie, Lu Jicang, Zhou Gang, et al. A review of research on document-level relationship extraction techniques[J]. Computer Science,2023,50(05):189-200.

[15] Yang Yu. Research on document-level relationship extraction technique[D]. Harbin Institute of Technology,2021.DOI:10.27061/d.cnki.ghgdu.2021.002868.