Document-Level Relation Extraction with Distance-dependent Bias Network and Neighbors Enhanced Loss

Hao Liang[1,2] [0000−0001−7629−7937] and Qifeng Zhou[1,2(✉)] [0000−0003−3583−6943]

[1] Department of Automation, Xiamen University, Xiamen 361005, China.
[2] Xiamen Key Laboratory of Big Data Intelligent Analysis and Decision-making, Xiamen 361005, China.
lianghao6@stu.xmu.edu.cn, zhouqf@xmu.edu.cn

Abstract. Document-level relation extraction (DocRE), in contrast to sentence-level, requires additional context to be considered. Recent studies, when extracting contextual information about entities, treat information about the whole document equally, which inevitably suffers from irrelevant information. This has been demonstrated to make the model not robust: it predicts correctly when an entire document is fed but errs when non-evidence sentences are removed. In this work, we propose three novel components to improve the robustness of the model by selectively considering the context of the entities. Firstly, we propose a new method for computing the distance between tokens that reduces the distance between evidence sentences and entities. Secondly, we add a distance-dependent bias network to each self-attention building block to exploit the distance information between tokens. Finally, we design an auxiliary loss for entities with higher attention to close tokens in the attention mechanism. Experimental results on three DocRE benchmark datasets show that our model not only outperforms existing models but also has strong robustness.

Keywords: Relation extraction · Self attention · Pre-training model.

1 Introduction

Relationship extraction (RE) aims to find predefined relations between entities from the texts. It is the fundamental of knowledge graph [2,38], question answering [7], and information extraction [17,28]. Early RE focused on the sentence-level [18,39], but realistic application scenarios often span across sentences, thus it is natural to shift to Document-level RE (DocRE) [31,43,32]. DocRE requires finding useful information from a large amount of context for logical reasoning, which is highly challenging.

Identifying the relation of an entity pair within a document by focusing on only a portion of the document is quite intuitive. Huang *et al.* [6] proposed a heuristic method to select three sentences for each pair of entities instead of inputting the whole document, which is effective but will delete a lot of helpful

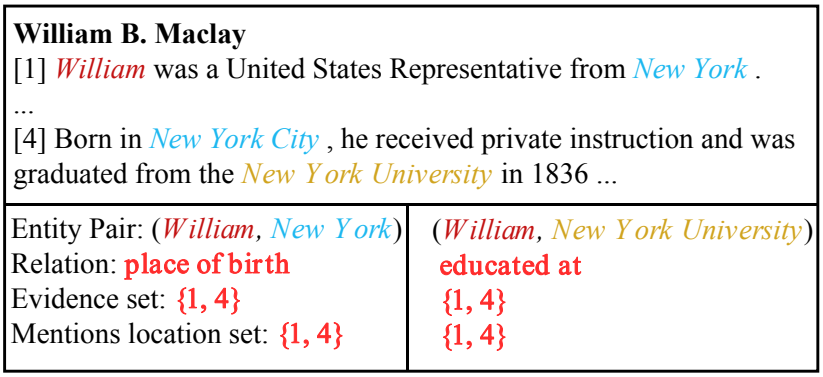

Fig. 1. An illustration on the DocRED dataset. New York and New York City are mention of the same entity.

contexts. Xu *et al.* [32] gives a weight to each sentence when identifying entity pairs, but the model requires two forward propagations to identify each pair of entities, which is not efficient. Nevertheless, these works point in a direction to improve the robustness and performance of the DocRE model, that is, the model should focus more on evidence sentences rather than non-evidence sentences.

The DocRED dataset [34] is labeled with the evidence sentences of an entity pair and all mentions of entities, Fig. 1 is an example. The entity pair (*William, New York*) has relation *a place of birth*. A human identifies this relation only by the sentences $\langle 1 \rangle$ and $\langle 4 \rangle$ (evidence sentences), that is, the Evidence set $\{1, 4\}$, denoted as set E. The Mentions location set is a set of sentences where all the mentions of this entity pair are located, denoted as set M. In this case, $E \subseteq M$, does this mean that we only need to consider the sentences in which the entities mention are located? In this regard, we made a statistic for DocRED dataset, see Table 1. **"0"** in the table means that for an entity pair, if we consider only the

Table 1. The percentage that $E \subseteq M$ as M expands in the DocRED dataset.

Expand number	0	1	2	3	4	5	6	7	...	13
Probability	90.9%	96.4%	98.1%	98.9%	99.3%	99.6%	99.7%	99.8%	...	100%

sentences in M, the percentage of $E \subseteq M$ is 90.9%. Now we extend one sentence outward with M as the center, that is, if M is $\{3\}$, it will be extended to $\{2, 3, 4\}$, then the percentage of $E \subseteq M$ is 96.4% as shown in **"1"** in the Table 1, and the rest of the values in the table have the same meaning in turn. From the Table 1, we may observe that the percentage growth is not significant as the number of considered sentences becomes larger. This represents that for an entity pair, the more distant token, the more noise and the less helpful information. Therefore, we assume that the DocRE model should pay differential attention to tokens at

different distances when identifying the relation of entity pairs, and the closer the token is, the more valuable it will be.

To this end, we propose **S**elf-Attention with **D**istance-dependent **B**ias **N**etwork (SDBN) and **N**eighbors **E**nhanced **L**oss (NEL). Specifically, we design a Bias Network that can improve the self-attention mechanism by using the mutual distance between tokens within a document as prior knowledge. Meanwhile, we design an auxiliary NEL to encourage the model to have a higher attention score for tokens that are closer to an entity pair. In addition, we use crossing-distance in our model, that is, the distance of an entity from other tokens is determined by the closest distance between the token and the entity mentions. Take Fig. 1 as an illustration, the distance between *New York City* in sentence $\langle 4 \rangle$ and *William* in sentence $\langle 1 \rangle$ is **3** (three sentences apart), but since *New York* in sentence $\langle 1 \rangle$ and *New York City* belong to the same entity, we define the crossing-distance between *New York City* and *William* as **0**, which can be seen as the context of *New York City* is enhanced from $\langle 4 \rangle$ to $\langle 1, 4 \rangle$.

2 Methodology

2.1 Crossing-Distance calculation

In identifying the relation of an entity pair, tokens with different distances have different bias network parameters to generate attention preferences. Thus we create an adjacency matrix to record the distance between tokens according to the following distance calculation.

Given a document $D = \{s_1, s_2, ..., s_N\}$, containing a set of entities $\{e_i\}_{i=1}^{n}$, where each sentence s_i is a sequence words. We convert D into a sequence of tokens $x = (x_1, x_2, ..., x_n)$, and a sequence $l = (l_1, l_2, ..., l_n), l_i \in \{1, 2, ..., N\}$ that records the sentence in which the token is located. Each entity may have multiple mentions, and each mention may consist of many tokens. Thus for each entity, we use a set $\left\{l_k^i\right\}_{k=1}^{N_{e_i}}$ to note in which sentences this entity is in. We classify all tokens into the following two categories.

- *Entity token*: Token that belongs to an entity.
- *Non-entity token*: Token that does not belong to any entity.

For set operations, we define the function $F(A, B)$, which works to determine the absolute minimum of the difference between the elements of A and B. There are three distance calculation methods derived between the two types of tokens:

Entity token and *Entity token*:

$$min(F(\left\{l_k^i\right\}_{k=1}^{N_{e_i}}, \left\{l_k^j\right\}_{k=1}^{N_{e_j}}), \mu) \tag{1}$$

where e_i and e_j represent the entities to which the two tokens belong, respectively, and μ is a hyper-parameter, the distance beyond μ is considered as μ. Shaw *et al.* [19] and Lee *et al.* [9] have done similar clipping in the calculation

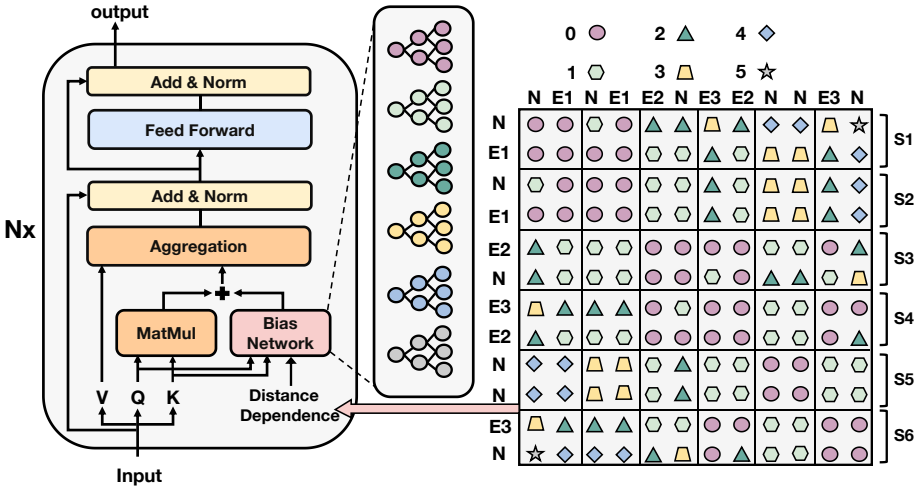

Fig. 2. The overall architecture of SDBN, when $\mu = 5$. Left illustrates structured self-attention as its basic building block. Right explains our bias network. This example consists of six sentences: S1, S2, S3, S4, S5, S6, and three entities: E1, E2, and E3. N denotes non-entity tokens. Element in row i and column j represents the distance of query token x_i to key token x_j, we use different graphic shapes to distinguish different distances.

of relative distances, which allows the model to be generalized to any sequence length.

Entity token and *Non-entity token*:

$$min(F(\{l_k^i\}_{k=1}^{N_{e_i}}, \{l_j\}), \mu) \tag{2}$$

Non-entity token and *Non-entity token*:

$$min(F(\{l_i\}, \{l_j\}), \mu) \tag{3}$$

For explanation, we assume that $\mu = 5$. Using (1)(2)(3), we construct the distance information of the entire document from each other as an adjacency matrix with elements from a finite set:$\{0, 1, 2, 3, 4, 5\}$.

2.2 SDBN

SDBN inherits the architecture of Transformer [23] encoder, which is a stack of the identical building block. As its core part, we added a bias network that can utilize the distance information in the self-attention mechanism. It makes the model generate entity representation with attention preferences for contexts of different distances.

A token sequence $x = (x_1, x_2, \ldots, x_n)$ is provided as input, following the calculate of 2.1, we introduce $A = \{a_{ij}\}$ to represent adjacency matrix, where

$i, j \in \{1, 2, \ldots, n\}$ and $a_{ij} \in \{0, 1, 2, 3, 4, 5\}$ is a discrete variable denotes the distance from x_i to x_j.

The input representation $x_i^m \in \mathbb{R}^{d_{in}}$ is first projected into the query/key/value vector respectively in each layer m.

$$\mathbf{q}_i^m = x_i^m W_m^Q, \mathbf{k}_i^m = x_i^m W_m^K, \mathbf{v}_i^m = x_i^m W_m^V \tag{4}$$

where $W_m^Q, W_m^K, W_m^V \in \mathbb{R}^{d_{in} \times d_{out}}$. Based on these inputs and the distance adjacency matrix A, we compute the raw attention scores and distance-dependent attention biases, then aggregate them together as the final attention scores to engage in the self-attention mechanism.

The raw attention score is produced by query-key product as in standard self-attention:

$$e_{ij}^m = \frac{\mathbf{q}_i^m \mathbf{k}_j^m}{\sqrt{d}} \tag{5}$$

To model the distance dependency based on their contextualized query / key representations, we use an additional module in parallel to it. We parameterize it as bias network that transforms a_{ij}, together with the query and key vectors $\mathbf{q}_i^m$ and $\mathbf{k}_j^m$, into an attentive bias, then apply it to e_{ij}^m:

$$\tilde{e}_{ij}^m = e_{ij}^m + \frac{bias\, network\left(\mathbf{q}_i^m, \mathbf{k}_j^m, a_{ij}\right)}{\sqrt{d}} \tag{6}$$

The proposed bias network regulates the attention flow from x_i to x_j. As a result, the model gains from the information provided by distance dependencies.

After obtaining the regulated attention scores $\tilde{e}_{ij}^m$, a softmax operation is used to aggregate the value vectors.

$$\mathbf{z}_i^{m+1} = \sum_{j=1}^{n} \frac{\exp \tilde{e}_{ij}^m}{\sum_{k=1}^{n} \exp \tilde{e}_{ik}^m} \mathbf{v}_j^m \tag{7}$$

here $\mathbf{z}_i^{m+1} \in \mathbb{R}^{d_{\text{out}}}$ is the updated contextual representation of x_i^m. Fig. 2 gives the overview of SDBN. In the next section, we describe the bias network.

2.3 Bias Network

We instantiate each discrete distance a_{ij} as neural layers with particular parameters, train, then apply them in a compositional manner to include them into an end-to-end trainable deep model. As a consequence, we get a structured model composed of corresponding layer parameters for each input adjacency matrix A made up of a_{ij}. Regarding the specific layout of these neural layers, we apply two options: Bilinear Transformation and Decomposed Linear Transformation.

$$\begin{aligned} \text{bias}_{ij}^m &= \text{Bilinear}\left(a_{ij}, \boldsymbol{q}_i^m, \boldsymbol{k}_j^m\right) \\ &\quad \text{or} \\ &= \text{Decomp}\left(a_{ij}, \boldsymbol{q}_i^m, \boldsymbol{k}_j^m\right) \end{aligned} \tag{8}$$

Bilinear Transformation Lin *et al.* [12] proposed bilinear to combine the features of two images, which simplifies the gradient calculation and can be directly applied to our bias network.

$$\text{bias}_{ij}^{m} = \mathbf{q}_i^m \mathbf{W}_{m,a_{ij}}^B \boldsymbol{k}_j^{m^T} + b_{m,a_{ij}} \tag{9}$$

here, a_{ij} is parameterized as a trainable neural layer $\mathbf{W}_{l,a_{ij}}^B \in \mathbb{R}^{d_{\text{out}} \times 1 \times d_{\text{out}}}$, which projects the query and key vector into a single-dimensional bias. Regarding the second term, we directly represent prior bias for each distance, regardless of its context, in $b_{m,a_{ij}}$.

Decomposed Linear Transformation We introduce bias on the query and key vector respectively, according to Xu *et al.* [31], thus the bias is decomposed into:

$$bias_{ij}^{l} = \mathbf{q}_i^m \mathbf{K}_{m,a_{ij}}^T + \mathbf{Q}_{m,a_{ij}} \mathbf{k}_j^m + b_{m,a_{ij}} \tag{10}$$

where $\mathbf{K}_{m,a_{ij}}, \mathbf{Q}_{m,a_{ij}} \in \mathbb{R}^d$ are also trainable neural layers.

So the overall computation of ditance-dependent self-attention is:

$$\begin{aligned} \tilde{e}_{ij}^m &= \frac{\mathbf{q}_i^m \mathbf{k}_j^{m^T} + bias\, network\left(\mathbf{q}_i^m, \mathbf{k}_j^m, a_{ij}\right)}{\sqrt{d}} \\ &= \frac{\mathbf{q}_i^m \mathbf{k}_j^{m^T} + \mathbf{q}_i^m \mathbf{W}_{m,a_{ij}}^B \mathbf{k}_j^m + b_{m,a_{ij}}}{\sqrt{d}} \\ or& \\ &= \frac{\mathbf{q}_i^m \mathbf{k}_j^{m^T} + \mathbf{q}_i^m \mathbf{K}_{m,a_{ij}}^T + \mathbf{Q}_{m,a_{ij}} \mathbf{K}_j^{m^T} + b_{m,a_{ij}}}{\sqrt{d}} \end{aligned} \tag{11}$$

Adaptive distance dependencies aren't shared between different layers or attention heads since Bias Network model them based on context.

2.4 Neighbors Enhanced Loss

The proposed SDBN model takes the document text D as input and builds its contextual representation within and throughout the encoding stage, guided by the distance. We create a fixed dimensional representation for each target entity using average pooling after the encoding stage. The entity representation is denoted as $e_i \in \mathbb{R}^{d_e}$. Thus for the ith entity pair in D, to determine whether there is a relation j, we have the following equation:

$$P_{ij} = sigmoid(\mathbf{e}_{ih} \mathbf{W}_j \mathbf{e}_{it}) \tag{12}$$

where $\mathbf{W}_j \in \mathbb{R}^{d_e \times d_e}$,$\mathbf{e}_{ih}$ is the head entity and $\mathbf{e}_{it}$ is the tail entity. The model is trained using a loss consisting of two components: Binary Cross-Entropy Loss and Neighbors Enhanced Loss.

Binary Cross-Entropy Loss This loss function is widely adopted and we also use it as part of the loss function.

$$\mathcal{L}_1 = -\sum_{i_h \neq i_t} \sum_{j \in \mathcal{R}} (r_{ij} \log P_{ij} + (1 - r_{ij}) \log(1 - P_{ij})) \tag{13}$$

where $\mathcal{R}$ denotes the set of relation types and $r_{ij} \in \{0, 1\}$ is the groundtruth label regarding entity pair i and relation j.

Neighbors Enhanced Loss Based on our findings in section 1, the closer the context is to an entity, the greater its value to that entity, especially after crossing-distance was used to enhance the entity context. To give the model this tendency, we designed Neighbors Enhanced loss. Since we use SDBN as an encoder, which has learned the crossing-distance dependencies at the token level, directly using their attention heads to enhance the attention to neighbors is appropriate.

Specifically, given a pre-trained multi-head attention matrix $\mathbf{G} \in \mathbb{R}^{H \times n \times n}$, where $\mathbf{G}_{kij}$ indicates attention score from token i to token j in the k^{th} attention head; n is the length of the input token sequence. Previously we have recorded the distance between token i and token j using a_{ij}. Suppose we want token i to pay more attention to contexts within the distance $\beta (\beta \leq \mu, \beta$ is a hyper-parameter.), we only need to use $a_{ij} \leq \beta$ to find the attention score of token i to these contexts. Therefore we use the following loss:

$$\mathcal{L}_2 = -log(\frac{\sum_{k \in H} \sum_{a_{ij} \leq \beta} \mathbf{G}}{\sum_{k \in H} \sum_{a_{ij} \in A} \mathbf{G}}) \tag{14}$$

$\mathcal{L}_2$ will push the attention score of $a_{ij} \leq \beta$ higher than the other attention scores. In our experiments, we use the attention matrix of the last SDBN layer, which avoids damaging the attentional flow and serves as a guide. Thus the total loss is:

$$\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_2 \tag{15}$$

where λ is a hyper-parameter to control the attention flow.

3 Experiment

3.1 Datasets

DocRED DocRED [34] is a relation extraction dataset created from Wikipedia and Wikidata. The dataset's documents are each annotated by humans with references to named entities, coreference data, intra- and inter-sentence relations, and supporting documentation. The dataset also offers vastly distantly supervised data in addition to human-annotated data. The two major metrics for evaluation are Ign F1 and F1 score according to Yao *et al.* [34], where Ign F1

is the F1 score that excludes triples from the annotated training data. The test results need to be submitted to the official Codalab.[1]

CDR CDR [11] is a human-annotated chemical-disease relation extraction dataset in biomedicine, consisting of 500 documents, which is tasked with predicting binary interactions between chemical and disease concepts.

GDA GDA [29] is a large-scale dataset in the biomedical domain and its task is to predict the binary interactions between genes and disease concepts. It contains 29192 documents for training, and we took 20% of the training set as the development set.

3.2 Implementation Details

Our model is implemented based on the Huggingface Transformers [27], using the pre-trained models Roberta-large [14] and Bert-large [4] and SciBERT-base [1]. For the DocRED dataset, we first train using the distantly supervised dataset to initialize the bias network. All hyper-parameters are grid searched and the best performers are used on the test set. The optimizer used for all experiments is Adam. The model is trained on a single NVIDIA V100 GPU with 32 GB memory.

3.3 Compared Methods

Due to the effectiveness of the pre-trained model, many models are constructed based on it, including ours. We mainly compare with these Transformer-based models.

- Wang *et al.* [26] built an enhanced Bert baseline using a two-stage prediction approach.
- Ye *et al.* [35] proposed a CorefBERT, which can help pre-trained models to better exploit co-reference relations in the context.
- Xu *et al.* [33] explicitly created paths such as logical reasoning and co-reference disambiguation in DocRE.
- Zhou *et al.* [43] proposed a method to enhance the entity context and introduce an adaptive threshold to solve the multi-label problem.
- Xu *et al.* [32] was designed with a loss function that allows the model to focus more on evidence sentences, which can reduce noise and improve the robustness of the model.
- Xie *et al.* [30] designed a lightweight model for extracting evidence sentences to be trained jointly with the RE model to help the RE model focus on evidence sentences.

[1] https://competitions.codalab.org/competitions/20717. Our model is named SDBN-DF.

– Zhang *et al.* [37] used the U-shaped Network in computer vision on DocRE to get global information at the entity-level.
– Xu *et al.* [31] used prior knowledge of entity structure to help model inference, which inspired our work.

Table 2. Results on DocRED. Subscript DL and BL refer to Decomposed Linear Transformation and Bilinear Transformation.

Models	Dev		Test	
	Ign F1	**F1**	**Ign F1**	**F1**
Coref-Roberta large [35]	57.35	59.43	57.90	60.25
BERT Two-stage [26]	-	54.42	-	53.92
GAIN+SIEF [32]	59.82	62.24	59.87	62.29
ATLOP-Roberta large [43]	61.46	63.37	61.39	64.40
DRN-Bert base [33]	59.33	61.39	59.15	61.37
EIDER-Roberta large [30]	62.48	64.37	62.85	64.79
DocuNet-Roberta large [37]	62.35	64.26	62.39	64.55
SSAN+Adaptation [31]	63.76	65.69	63.78	65.92
SDBN_{DL}+Bert large	63.32	65.38	63.38	65.42
SDBN_{BL}+Bert large	63.47	65.34	63.58	65.46
SDBN_{DL}+Roberta large	64.38	66.92	64.62	67.03
SDBN_{BL}+Roberta large	**64.72**	**67.26**	**64.88**	**67.12**

3.4 Main Results

Table 2 shows the main results on the DocRED dataset. We used Bert large and Roberta large pre-trained models on *DF* and *BL*, respectively. Similar to the results of other papers, Roberta large gives a significantly better result than Bert large. *BL* gives a slightly better result than *DL*, indicating that the former is better adapted to the introduced crossing-distance.

We compared the model EIDER-Roberta large [30], which also utilizes evidence sentences, and obtained a 2.03/2.33 boost on the test set of lgnF1/F1. Meanwhile, we compared SSAN [31], which also introduces attention bias, and obtained a 1.1/1.2 improvement on lgnF1/F1 of the test set. The comparison with other models demonstrates the effectiveness of our SDBN. Table 3 shows the main results of CDR and GDA datasets. We use the SciBERT base as a pre-trained model, which has a better performance in the biomedical domain. On the CDR and GDA test set, we improved by 2.10/1.72 over CGM2IR-SciBERT [41] on F1, which is already outperforming all existing work. These results demonstrate that our approach is highly applicable and generalizable.

3.5 Ablation Study

On DocRED, we conducted an ablation study of the best-performing method, SDBN_{BL}+Roberta large, by turning off one component at a time. The results

of the model ablation study are shown in Table 4. It is obvious that the results of the model decrease significantly when there is no Bias Network in the model. However, when the NEL is removed, the Bias Network alone also gets better results. It implies that our proposed crossing-distance is a better guide to the attention flow and the model benefits from it.

The NEL enhancement is not apparent, because the context at close range, although helpful, also contains distracting information. However, NEL can influence the tendency of attention bias in the Bias Network, and the two complement each other to achieve better results.

Table 3. Results on CDR and GDA.

Model	CDR	GDA
ATLOP-SciBERT [43]	69.4	83.90
EIDER-SciBERT [30]	70.6	84.54
CGM2IR-SciBERT [41]	73.8	84.70
SSAN-SciBERT [31]	68.7	83.70
SDBN$_{DL}$+SciBERT base	**75.9**	86.08
SDBN$_{BL}$+SciBERT base	75.2	**86.42**

Table 4. Ablation Study of SDBN on DocRED. We turn off different components of the model one at a time.

Model	Ign F1	F1
SDBN$_{BL}$+Roberta large	64.88	67.12
- Bias Network	61.82	63.38
- Neighbors Enhanced Loss	63.62	66.24
- both	59.47	61.42

3.6 Robustness Analysis

We designed an experiment to verify the robustness of our model. In the training stage, we make no changes and input the whole document. During the testing stage, we randomly remove two of the non-evidence sentences that are not evidence of any entities from each document in the development set. Every 5 epochs, the result of F1 is compared to the performance of the model without deleting sentences, and the absolute value of the difference is recorded.

Our approach is compared with Xie *et al.* [30] and Xu *et al.* [32], both of which share the central notion of making the model concentrate more on the evidence sentences. The comparison results are shown in Figure 3. Obviously, our SDBN changes the least after randomly deleting two non-evidence sentences, since this does not change the crossing-distance of the evidence sentences from the entity. This shows the robustness of our model, as it is less sensitive to non-evidence sentences for DocRE.

3.7 Hyper-parameter Analysis

Our model has two hyper-parameters μ and β. μ is a maximum distance cut when computing the crossing-distance, and distance exceeding μ will be cut to μ. The meaning of β is that the token needs to be more focused on the context within β distance.

The best result of our grid search is $\mu = 5$ and $\beta = 3$. Therefore, our experiments on F1 score fix one of the hyperparameters and then change the other one. The experimental results are in Figure 4. Since $\beta \leq \mu$, the part of $\beta > \mu$

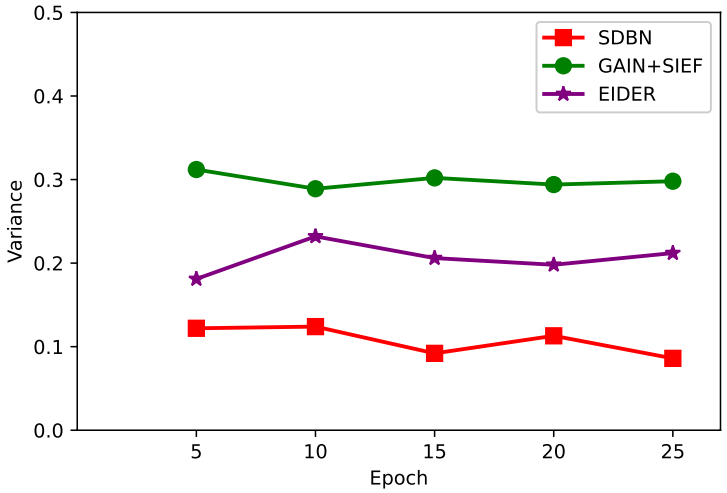

Fig. 3. Robustness of DocRE models.

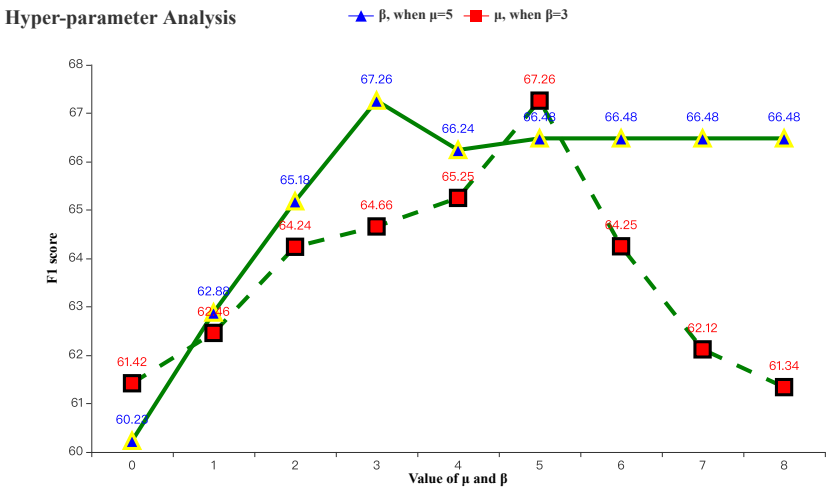

Fig. 4. Performances of the classification (in F1 score) on the development set of different hyper-parameter β and μ.

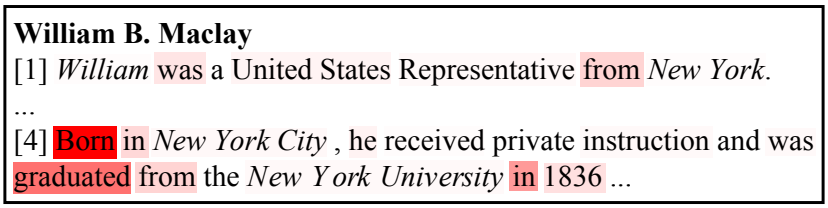

Fig. 5. Case study on DocRED dataset.

in Figure 4 we replace with the value of $\beta = \mu$. As seen, the performance of the model starts to degrade for $\mu > 5$, which is because the larger μ is, the worse the ability of our model to scale to arbitrary sequence lengths.

3.8 Case Study

We visualize the attention score of *William* to other tokens when identifying the relation of an entity pair (*William, New York*) in the example of Fig. 1. As shown in Fig. 5, *William* gives high weight to *Born* and *graduated*. *Born* is evidence for the relation *place of birth*. The reason why the weight of *graduated* is also high is perhaps due to the similarity of *New York* and *New York University* in terms of word embedding. The visualization demonstrates that our proposed three components do not compromise the attention mechanism. Entities are still able to pay attention to the evidence properly even if they are several sentences apart.

4 Related Work

DocRE models can be broadly classified into the following three categories:

Sequence-based Models These models encode the whole document using neural architectures like CNN [13,16] and bidirectional LSTM [21,20], then obtain entity embeddings and predict relations for each entity pair utilizing bilinear function. Such sequence models do not work very effectively for modeling complex contexts and are relatively obsolete work.

Graph-based Models These models construct graphs based on the mention, sentence, and paragraph of the document and employ a variety of graph networks for inference [25,15,5,3,36,10,42,40,8,24]. The essence of Graph Convolutional Networks (GCN) is to learn the manner of aggregating neighbors, and to some extent can learn the way of aggregation at different distances. But this aggregation depends on the construction method and depth of the graph network, and will not personalize the aggregation in various ways according to different nodes. The Graph Attention Networks (GAT) will be personalized to calculate

aggregation weights based on node information, but the distance information is weakened. Our work can be seen as a neutralization of GCN and GAT, calculating the attention of one token to another token considering both the personalized embedding of the two as well as the distance between them.

Transformer-based Models Without using graph structures, these models adapt pre-trained language models directly to DocRE [22,35,26,43,33,37]. These models achieve great performance based on the strong adaptability of the pre-trained models.

Xu *et al.* [31] used a priori information on entity structure to guide the attention flow of the pre-trained model, which inspired our work. Unlike it, we propose a new distance calculation method to guide the attention flow and design an auxiliary loss to help the model make better utilization of the attention bias.

5 Conclusion

In this work, we propose three novel techniques SDBN and NEL, as well as a new way of computing the distance between tokens. The new distance enhances the context of entities; SDBN allows tokens to have different attention preferences for contexts of different distances; NFL allows tokens to pay more attention to closer contexts. The three can be perfectly combined to enable the model to pay more attention to useful contexts. This will reduce the interference of irrelevant information when identifying the relation of entity pairs.

Pronouns do not belong to any entity, hence they cannot benefit from crossing-distance. The model can only rely on the powerful encoding capabilities of self-attention to implicitly exploit pronoun information. In the future, we intend to refer to Ye *et al.* [35], which explicitly exploits pronoun information.

Acknowledgement. We would like to thank the anonymous reviewers for their insightful feedback and comments. This work is partially supported by China Natural Science Foundation under grant (No. 62171391) and the Natural Science Foundation of Fujian Province of China under grant (No. 2020J01053).

References

1. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3615–3620. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1371, https://aclanthology.org/D19-1371
2. Cao, Y., Ji, X., Lv, X., Li, J., Wen, Y., Zhang, H.: Are missing links predictable? an inferential benchmark for knowledge graph completion. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the

11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 6855–6865. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.534, https://aclanthology.org/2021.acl-long.534
3. Christopoulou, F., Miwa, M., Ananiadou, S.: Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4925–4936. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1498, https://aclanthology.org/D19-1498
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018), https://arxiv.org/abs/1810.04805
5. Guo, Z., Zhang, Y., Lu, W.: Attention guided graph convolutional networks for relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 241–251. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1024, https://aclanthology.org/P19-1024
6. Huang, Q., Zhu, S., Feng, Y., Ye, Y., Lai, Y., Zhao, D.: Three sentences are all you need: Local path enhanced document relation extraction. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 998–1004. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-short.126, https://aclanthology.org/2021.acl-short.126
7. Jia, R., Lewis, M., Zettlemoyer, L.: Question answering infused pre-training of general-purpose contextualized representations. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 711–728. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.59, https://aclanthology.org/2022.findings-acl.59
8. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016), https://arxiv.org/abs/1609.02907
9. Lee, B.K., Lessler, J., Stuart, E.A.: Weight trimming and propensity score weighting. PloS one **6**(3), e18174 (2011)
10. Li, B., Ye, W., Sheng, Z., Xie, R., Xi, X., Zhang, S.: Graph enhanced dual attention network for document-level relation extraction. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 1551–1560. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.136, https://aclanthology.org/2020.coling-main.136
11. Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: Biocreative v cdr task corpus: a resource for chemical disease relation extraction. Database **2016** (2016)
12. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 1449–1457 (2015)
13. Liu, C., Sun, W., Chao, W., Che, W.: Convolution neural network for relation extraction. In: International conference on advanced data mining and applications. pp. 231–242. Springer (2013)

14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019), https://arxiv.org/abs/1907.11692
15. Nan, G., Guo, Z., Sekulic, I., Lu, W.: Reasoning with latent structure refinement for document-level relation extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1546–1557. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.141, https://aclanthology.org/2020.acl-main.141
16. Nguyen, T.H., Grishman, R.: Relation extraction: Perspective from convolutional neural networks. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. pp. 39–48. Association for Computational Linguistics, Denver, Colorado (Jun 2015). https://doi.org/10.3115/v1/W15-1506, https://aclanthology.org/W15-1506
17. Papanikolaou, Y., Staib, M., Grace, J.J., Bennett, F.: Slot filling for biomedical information extraction. In: Proceedings of the 21st Workshop on Biomedical Language Processing. pp. 82–90. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.bionlp-1.7, https://aclanthology.org/2022.bionlp-1.7
18. Park, S., Kim, H.: Improving sentence-level relation extraction through curriculum learning. arXiv preprint arXiv:2107.09332 (2021), https://arxiv.org/abs/2107.09332
19. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 464–468. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2074, https://aclanthology.org/N18-2074
20. Song, L., Zhang, Y., Wang, Z., Gildea, D.: N-ary relation extraction using graph-state LSTM. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2226–2235. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1246, https://aclanthology.org/D18-1246
21. Sorokin, D., Gurevych, I.: Context-aware representations for knowledge base relation extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1784–1789. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1188, https://aclanthology.org/D17-1188
22. Tang, H., Cao, Y., Zhang, Z., Cao, J., Fang, F., Wang, S., Yin, P.: Hin: Hierarchical inference network for document-level relation extraction. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 197–209. Springer (2020)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
24. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017), https://arxiv.org/abs/1710.10903
25. Wang, D., Hu, W., Cao, E., Sun, W.: Global-to-local neural networks for document-level relation extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 3711–3721. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.303, https://aclanthology.org/2020.emnlp-main.303

26. Wang, H., Focke, C., Sylvester, R., Mishra, N., Wang, W.: Fine-tune bert for docred with two-step process. arXiv preprint arXiv:1909.11898 (2019), https://arxiv.org/abs/1909.11898
27. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6
28. Wu, X., Zhang, J., Li, H.: Text-to-table: A new way of information extraction. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2518–2533. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.180, https://aclanthology.org/2022.acl-long.180
29. Wu, Y., Luo, R., Leung, H., Ting, H.F., Lam, T.W.: Renet: A deep learning approach for extracting gene-disease associations from literature. In: International Conference on Research in Computational Molecular Biology. pp. 272–284. Springer (2019)
30. Xie, Y., Shen, J., Li, S., Mao, Y., Han, J.: Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 257–268. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.23, https://aclanthology.org/2022.findings-acl.23
31. Xu, B., Wang, Q., Lyu, Y., Zhu, Y., Mao, Z.: Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14149–14157 (2021), https://arxiv.org/abs/2102.10249
32. Xu, W., Chen, K., Mou, L., Zhao, T.: Document-level relation extraction with sentences importance estimation and focusing. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2920–2929. Association for Computational Linguistics, Seattle, United States (Jul 2022), https://aclanthology.org/2022.naacl-main.212
33. Xu, W., Chen, K., Zhao, T.: Discriminative reasoning for document-level relation extraction. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 1653–1663. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.144
34. Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., Sun, M.: DocRED: A large-scale document-level relation extraction dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 764–777. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1074, https://aclanthology.org/P19-1074
35. Ye, D., Lin, Y., Du, J., Liu, Z., Li, P., Sun, M., Liu, Z.: Coreferential Reasoning Learning for Language Representation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7170–7186. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.582, https://aclanthology.org/2020.emnlp-main.582

36. Zeng, S., Xu, R., Chang, B., Li, L.: Double graph based reasoning for document-level relation extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1630–1640. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.127, https://aclanthology.org/2020.emnlp-main.127
37. Zhang, N., Chen, X., Xie, X., Deng, S., Tan, C., Chen, M., Huang, F., Si, L., Chen, H.: Document-level relation extraction as semantic segmentation. arXiv preprint arXiv:2106.03618 (2021), https://www.ijcai.org/proceedings/2021/0551.pdf
38. Zhang, Y., Li, P., Liang, H., Jatowt, A., Yang, Z.: Fact-tree reasoning for n-ary question answering over knowledge graphs. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 788–802. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.66, https://aclanthology.org/2022.findings-acl.66
39. Zhang, Y., Qi, P., Manning, C.D.: Graph convolution over pruned dependency trees improves relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2205–2215. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1244, https://aclanthology.org/D18-1244
40. Zhang, Z., Yu, B., Shu, X., Liu, T., Tang, H., Yubin, W., Guo, L.: Document-level relation extraction with dual-tier heterogeneous graph. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 1630–1641. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.143, https://aclanthology.org/2020.coling-main.143
41. Zhao, C., Zeng, D., Xu, L., Dai, J.: Document-level relation extraction with context guided mention integration and inter-pair reasoning. arXiv preprint arXiv:2201.04826 (2022)
42. Zhou, H., Xu, Y., Yao, W., Liu, Z., Lang, C., Jiang, H.: Global context-enhanced graph convolutional networks for document-level relation extraction. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 5259–5270. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.461, https://aclanthology.org/2020.coling-main.461
43. Zhou, W., Huang, K., Ma, T., Huang, J.: Document-level relation extraction with adaptive thresholding and localized context pooling. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 14612–14620 (2021), https://arxiv.org/abs/2010.11304