SRF: Enhancing Document-Level Relation Extraction with a Novel Secondary Reasoning Framework

Fu Zhang[†], Qi Miao[†], Jingwei Cheng[*], Hongsen Yu, Yi Yan, Xin Li, Yongxue Wu
School of Computer Science and Engineering, Northeastern University, China
`{zhangfu,chengjingwei}@mail.neu.edu.cn; miaoqi02@baidu.com`

Abstract

Document-level Relation Extraction (DocRE) aims to extract relations between entity pairs in a document and poses many challenges as it involves multiple mentions of entities and cross-sentence inference. However, several aspects that are important for DocRE have not been considered and explored. Existing work ignore bidirectional mention interaction when generating relational features for entity pairs. Also, sophisticated neural networks are typically designed for cross-sentence evidence extraction to further enhance DocRE. More interestingly, we reveal a noteworthy finding: If a model has predicted a relation between an entity and other entities, this relation information may help infer and predict more relations between the entity's adjacent entities and these other entities. Nonetheless, none of existing methods leverage secondary reasoning to exploit results of relation prediction. To this end, we propose a novel **S**econdary **R**easoning **F**ramework (**SRF**) for DocRE. In SRF, we initially propose a DocRE model that incorporates *bidirectional mention fusion* and a simple yet effective *evidence extraction* module (incurring only an additional learnable parameter overhead) for relation prediction. Further, for the first time, we elaborately design and propose a novel *secondary reasoning* method to discover more relations by exploring the results of the first relation prediction. Extensive experiments show that SRF achieves SOTA performance and our secondary reasoning method is both effective and general when integrated into existing models.[1]

1 Introduction

Relation extraction (RE), which aims to identify semantic relations between head-tail entity pairs in a single sentence, is one of the most fundamental tasks in information extraction (Zeng et al., 2015; Zhang et al., 2018; Soares et al., 2019). However, in the real world, many relations are inferred through multiple sentences (Verga et al., 2018; Yao et al., 2019), which is referred to as *document-level relation extraction* (DocRE). This requires a model that can capture complex interactions between entities throughout the entire document.

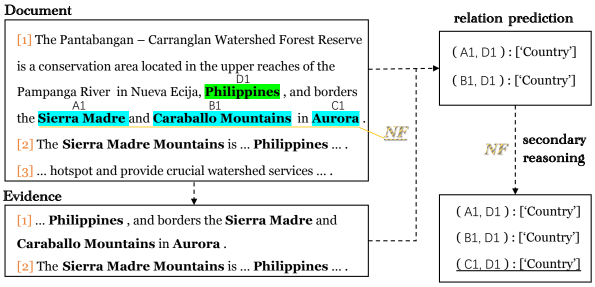

Document
[1] The Pantabangan – Carranglan Watershed Forest Reserve is a conservation area located in the upper reaches of the Pampanga River in Nueva Ecija, Philippines, and borders the Sierra Madre and Caraballo Mountains in Aurora .
D1
A1
B1
C1
[2] The Sierra Madre Mountains is ... Philippines ... .
[3] ... hotspot and provide crucial watershed services ... .
NF
Evidence
[1] ... Philippines , and borders the Sierra Madre and Caraballo Mountains in Aurora .
[2] The Sierra Madre Mountains is ... Philippines ... .
relation prediction
(A1, D1) : ['Country']
(B1, D1) : ['Country']
NF
secondary reasoning
(A1, D1) : ['Country']
(B1, D1) : ['Country']
(C1, D1) : ['Country']

Figure 1: A simple example of DocRE and a rough illustration of our idea of secondary reasoning. NF refers to Noun Fragment as will be defined in Section 2.

DocRE presents unique challenges compared to sentence-level RE. In DocRE, an entity pair may appear in multiple sentences and each entity may have multiple mentions (e.g., Sierra Madre Mountains is a mention of Caraballo Mountains as shown in Figure 1), which require models to accurately identify relevant contexts for inference. Moreover, an entity pair may have multiple types of relations in DocRE, whereas in sentence-level RE, an entity pair may have only one type of relation. This makes DocRE more challenging. Additionally, from statistics of widely used DocRE datasets DocRED (Yao et al., 2019) and its revised version Re-DocRED (Tan et al., 2022b), the majority of relation instances (61.1%) require reasoning to be identified, highlighting the importance of reasoning in DocRE.

To capture complex correlations between entity-level, mention-level, and sentence-level, some graph-based and non-graph-based models are proposed. Graph-based models mainly employ graph neural networks (Kipf and Welling, 2016) to per-

[1] Code is available at `https://github.com/zelf0914/SRF`.
[†] Equal contribution. [*] Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15426–15439
November 12-16, 2024 ©2024 Association for Computational Linguistics

form explicit reasoning on constructed document-level graphs, e.g., GAIN (Zeng et al., 2020) and SSAN (Xu et al., 2021a). Considering that Transformer (Vaswani et al., 2017) can implicitly model long-range dependencies, many non-graph-based models perform implicit reasoning using pre-trained models, including ATLOP (Zhou et al., 2021) and Eider (Xie et al., 2022) (please refer to Related Work in **Appendix A.1** for details).

However, when computing relation representation of an entity pair, the existing work ignores the bidirectional interaction between multiple mentions of the head and tail entities. *Moreover*, some methods (e.g., Eider (Xie et al., 2022)) attempt to enhance relation extraction performance by incorporating a subtask of evidence extraction. This subtask is designed to find evidence sentences from the document for each entity pair using a multi-task strategy. However, the existing methods necessitate the design of specialized neural networks to extract evidence sentences, which may result in significant additional overhead. *Additionally*, it is particularly noteworthy that none of the methods utilize secondary reasoning on relation prediction results. While most existing models can accurately predict relations between entity pairs with rich contextual connections, they struggle to predict rarely mentioned entities (RMEs), e.g., the entity Aurora in Figure 1. Such RMEs may be predicted depending on adjacent entities. If a model has predicted relations between these adjacent entities and other entities, this relation information can help infer and uncover more relations between the RMEs and these other entities, as we briefly illustrate in Figure 1 and will go into detail in Section 3.3. Therefore, it is particularly important to re-explore relation prediction results and perform secondary reasoning, as this is likely to predict more relations.

To address these challenges, we propose a novel **S**econdary **R**easoning **F**ramework (**SRF**) for DocRE, which includes several main contributions:

- We propose a bidirectional attention and fusion method based on entity-level and mention-level features of head-tail entities, which can better learn relational features of the entity pairs.

- We propose a simple yet effective evidence extraction module based on evidence words, by skillfully utilizing mention-level and entity-level relation attention scores. Instead of designing a specialized evidence extraction neural network as existing methods do, our method only incurs overhead of an additional learnable parameter.

- For the first time, we innovatively propose a secondary reasoning module that is carefully designed to further mine and utilize the results of the first relation prediction, with the aim of predicting more relations to improve performance.

- We conduct extensive experiments on the widely used datasets DocRED (Yao et al., 2019) and Re-DocRED (Tan et al., 2022b). Results demonstrate the superiority of our model. Further analyses demonstrate that our secondary reasoning module is both effective and general when integrated into other existing models.

2 Problem Formulation

Given a document D, DocRE is to predict a subset of relations from $R \cup \{NA\}$ between entity pairs (e_h, e_t), where *NA* indicates no relation. An entity e_μ can occur multiple times by its entity *mentions* $\{m_{\mu i}\}_{i=1}^{N_{e_\mu}}$ (where N_{e_μ} denotes the number of mentions of e_μ). A relation exists between (e_h, e_t) if it is expressed by any pair of their mentions.

We introduce a new term, **Noun Fragment (NF)**, which refers to a fragment that has at least 3 entities and must start with an entity and end with an entity as shown in Figure 1. There cannot be verbs and prepositions (except "of", "in") between the start and end entities. We allow "of" or "in" in NFs, e.g., "Duchy of Lorraine". Also, cases like "data mining" where "mining" is a verb yet the overall term functions as a part of an NF. These NFs can be extracted from datasets by the algorithm in **Appendix A.2**.

3 Methodology

An illustration of our SRF is shown in **Figure 2**, which consists of several main components:

(*i*) The *relation extraction module* (Section 3.1) captures relational features of an entity pair at entity-level and mention-level based on our bidirectional attention and fusion method. The encoding layer is introduced within this module.

(*ii*) The *evidence extraction module* (Section 3.2) extracts evidence words by introducing only a learnable parameter as overhead. During training, we *jointly train the relation and evidence extraction modules*, which have their own classifiers and share a base encoder. After the training, the relation extraction module will perform relation prediction.

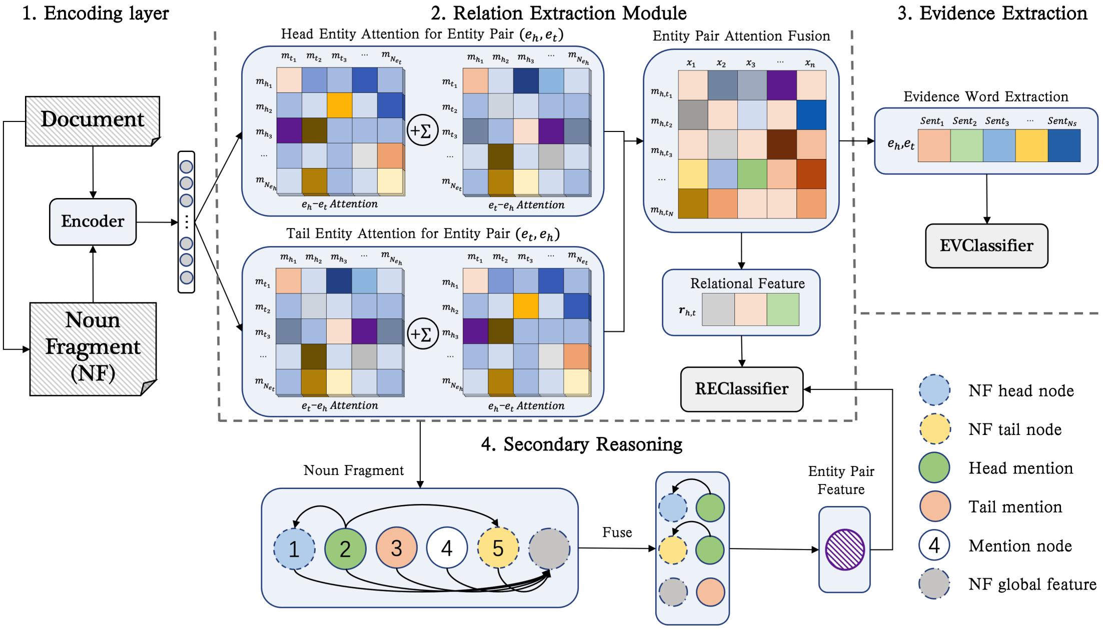


Figure 2: The overall architecture of our SRF for DocRE.

(*iii*) The *secondary reasoning module* (Section 3.3) is performed based on the above relation prediction results. The relations predicted from the relation extraction and secondary reasoning modules are *used together as final prediction results*.

3.1 Relation Extraction Module

The relation extraction module is designed to extract the relation between a given entity pair (e_h, e_t). We propose a *bidirectional attention and fusion* method based on head-tail entities, which aims to better learn relational features of entity pairs. Specifically, the attention score of the head entity e_h is composed of both the head-to-tail attention score and the tail-to-head attention score, with learnable parameters. When the tail entity e_t pays attention to the head entity e_h, this tail-to-head attention score increases the overall attention score of the head entity e_h. Being different from conventional methods that average 12 layers of multi-head attention at encoding layer, we use the top-k layers.

Given a document D, the encoding part first sends a document containing n words $x_1, x_2, ..., x_n$ into an encoder, such as XLNET (Yang et al., 2019), to obtain the representation of each word:

$$\boldsymbol{M} = \text{Encoder}([x_1, x_2, ..., x_n]) \qquad (1)$$

where $\boldsymbol{M} \in \mathbb{R}^{n \times d}$ and d is the hidden dimension.

Then, for an entity pair (e_h, e_t), we first obtain features of e_h and e_t. The weight of each mention m_{hi} of the head entity e_h is calculated following **three steps**: (**i**) Obtain the attention score of m_{hi} to the tail entity e_t, by fusing the importance scores from m_{hi} to each mention m_{tj} of the tail entity e_t.

$$\boldsymbol{a}_{m_{hi}} = \log \sum_{j=1}^{N_{e_t}} \exp \mathbf{A}_{m_{hi},m_{tj}} \qquad (2)$$

where $\mathbf{A}_{m_{hi},m_{tj}}$ is defined as follows:

$$\mathbf{A}_{x_i,x_j} = \frac{1}{k} \sum_{f=rank(1)}^{rank(k)} \boldsymbol{att}\left(x_i, x_j, f\right) \qquad (3)$$

where $\boldsymbol{att}\left(x_i, x_j, f\right) \in \mathbb{R}^1$ denotes attention score ranked f among multi-head attention scores between x_i and x_j in encoding layer. (**ii**) Similarly, obtain the attention score $\boldsymbol{a}_{m_{tj}}$ of each mention m_{tj} of tail entity e_t to the head entity e_h. (**iii**) Fuse the above attention scores to obtain the weight of each mention m_{hi} of the head entity e_h.

$$\boldsymbol{W}'_{m_{hi}} = \boldsymbol{a}_{m_{hi}} + \boldsymbol{W} \frac{1}{N_{e_t}} \sum_{j=1}^{N_{e_t}} \boldsymbol{a}_{m_{tj}} \qquad (4)$$

$$\boldsymbol{W}_{m_{hi}} = \frac{\exp(\boldsymbol{W}'_{m_{hi}})}{\sum_{c=1}^{N_{e_h}} \exp(\boldsymbol{W}'_{m_{hc}})} \qquad (5)$$

where $\boldsymbol{W} \in \mathbb{R}^1$ is a learnable parameter.

The *head entity feature* is obtained by fusing entity-level features and mention-level features.

We use two feature fusion strategies: concatenation strategy (Eq.6) and summation strategy (Eq.7).

$$\boldsymbol{e}_h = \left[\sum_{i=1}^{N_{e_h}} \boldsymbol{W}_{m_{hi}} \cdot \boldsymbol{M}_{m_{hi}} ; \sum_{i=1}^{N_{e_h}} \frac{\boldsymbol{M}_{m_{hi}}}{N_{e_h}} \right] \boldsymbol{W}_1 \tag{6}$$

$$\begin{aligned} \boldsymbol{e}_h = \boldsymbol{W}_2 \sum_{i=1}^{N_{e_h}} \boldsymbol{W}_{m_{hi}} \cdot \boldsymbol{M}_{m_{hi}} \\ +(1-\boldsymbol{W}_2)\frac{1}{N_{e_h}} \sum_{i=1}^{N_{e_h}} \boldsymbol{M}_{m_{hi}} \end{aligned} \tag{7}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{2d \times d}$, $\boldsymbol{W}_2 \in \mathbb{R}^1$ are learnable parameters. $\boldsymbol{M}_{m_{hi}}$ is the mention feature obtained by Eq.(1), where the entity mentions are marked by a token "*" at the start and end position, and we take the embedding of the token "*" at the start of the mention as its embedding. $[;]$ denotes concatenation. The calculation process of *tail entity feature* is the same as that of the head entity.

Simultaneously, we determine the weight of the head entity e_h for the document D, by utilizing the weight $\boldsymbol{W}_{m_{hi}}$ of each mention of the head entity and the importance score $\mathbf{A}_{m_{hi},x_i}$ of each mention for every word x_i in the document (similar to weight $\boldsymbol{W}_{tD}$ of the tail entity e_t).

$$\boldsymbol{W}_{hD} = \sum_{i=1}^{N_{e_h}} \frac{\boldsymbol{W}_{m_{hi}}}{N_{e_h}} \left[\mathbf{A}_{m_{hi},x_1}, ..., \mathbf{A}_{m_{hi},x_n} \right] \tag{8}$$

Now, we obtain the *relational feature* of (e_h, e_t).

$$\boldsymbol{R}'_{h,t} = \boldsymbol{W}_{hD} \cdot \boldsymbol{W}_{tD} \tag{9}$$

$$\boldsymbol{r}_{h,t} = \boldsymbol{R}'_{h,t} \times \boldsymbol{M} \tag{10}$$

Finally, we get relation prediction scores:

$$\mathbb{P}'_r = \text{FFNN}\left([\boldsymbol{e}_h ; \boldsymbol{r}_{h,t}]^T \boldsymbol{W}_r \, [\boldsymbol{e}_t ; \boldsymbol{r}_{h,t}] + \boldsymbol{b}_r \right) \tag{11}$$

where $\boldsymbol{W}_r \in \mathbb{R}^{2d \times 2d}$ and $\boldsymbol{b}_r \in \mathbb{R}$ are learnable parameters.

3.2 Evidence Extraction

In this module, we propose a simple and novel evidence extraction strategy that aims to find evidences for each entity pair from the document. Existing methods (e.g., Eider (Xie et al., 2022)) require specialized neural networks to extract evidence sentences, while our method only requires a learnable parameter overhead.

When extracting evidences for entity pairs, we find that, in the weight of an entity pair for the document, the sentence containing words x_i with high weights (*evidence words*) is often an evidence sentence. As shown in Eq.(8), when calculating the weight of an entity for a document, we use the method of *weighting* entity mentions, which can well consider the relationships between entity mentions. However, due to the excessive influence of some mentions on the weight, the impact of other mentions on the document may be ignored. Thus, in the evidence extraction module, we introduce another method of *averaging* entity mentions for calculating the weight of an entity pair for a document. On this basis, we adopt a learnable parameter to fuse the weights obtained by the two methods.

The method of *averaging* entity mentions for calculating the weight of an entity pair (e_h, e_t) for a document D is defined as follows:

$$\boldsymbol{R}''_{h,t} = \boldsymbol{W}'_{hD} \cdot \boldsymbol{W}'_{tD} \tag{12}$$

where the weight $\boldsymbol{W}'_{hD}$ of the head entity e_h for the document D (similar for $\boldsymbol{W}'_{tD}$) can be obtained:

$$\boldsymbol{W}'_{hD} = \frac{1}{N_{e_h}} \sum_{i=1}^{N_{e_h}} \left[\mathbf{A}_{m_{hi},x_1}, ..., \mathbf{A}_{m_{hi},x_n} \right] \tag{13}$$

We further adopt a *learnable parameter* $\boldsymbol{W}_{evi} \in \mathbb{R}^1$ to fuse the weights obtained by Eq.(9) and (12).

$$\boldsymbol{R}_{h,t} = \boldsymbol{W}_{evi} \boldsymbol{R}'_{h,t} + (1 - \boldsymbol{W}_{evi}) \boldsymbol{R}''_{h,t} \tag{14}$$

After obtaining the weight of the entity pair for the document, we further find the word with the highest weight in each sentence in the document. This word will be used as the *evidence word* of each sentence, and then the corresponding *evidence word vector* of the entity pair will be obtained.

$$\boldsymbol{S}_{evi}^{words} = [\max(\boldsymbol{R}_{h,t}^{sent1}), ... \max(\boldsymbol{R}_{h,t}^{sentN})] \tag{15}$$

Considering that values in evidence word vectors are generally between 0 and 0.2, so we use normalization to scale the values. The normalized vector is used directly as the prediction scores for evidence sentences of the entity pair.

$$\boldsymbol{S}_{evi} = \frac{\boldsymbol{S}_{evi}^{words}}{\max(\boldsymbol{S}_{evi}^{words}) + \min(\boldsymbol{S}_{evi}^{words})} \tag{16}$$

To reduce the noise generated by non-evidence sentences in model prediction, we set a learnable parameter η based on the results of the DocRED dev set. For each entity pair, if the score of a word in the corresponding evidence word vector exceeds η, the sentence where the word is located is considered as an evidence sentence. Finally, we concatenate evidence sentences to construct pseudo-documents (Xie et al., 2022) for prediction.

3.3 Secondary Reasoning

Secondary reasoning aims to further infer and mine more relations in the document based on the relation extraction results of Section 3.1 (i.e., the first predicted results). To achieve this, we explore the relations between entities within *Noun Fragments* (*NFs*, defined in Section 2) and propose a secondary reasoning method to mine more relations.

Based on the observation that existing models can correctly predict some entities within NFs based on context, favoring those with rich contextual connections. However, rarely mentioned entities in NFs might be predicted through their adjacent entities. When some entities are predicted to have relations with others, their adjacent entities are likely to share those relations. Therefore, re-examining prediction results and performing secondary reasoning is crucial. Note that, our secondary reasoning differs fundamentally from previous work like GAIN (Zeng et al., 2020) and DocRE-II (Zhang et al., 2022), which make a single prediction based on entity and relation representations. In contrast, our approach conducts a second round of reasoning to uncover additional relations after the initial prediction.

Specially, our secondary reasoning consists of *mention feature extraction*, *Noun Fragment feature extraction*, and *mention reasoning*.

3.3.1 Mention features

Assuming that the relation $r \in R$ between the entity pair (e_h, e_t) has been predicted through Section 3.1, the goal of secondary reasoning is to further predict whether there is the relation r between an entity pair $(e_h^{'}, e_t)$, where $e_h^{'}$ is an entity in a corresponding NF of e_h that was not predicted to have the relation r (that is, $e_h^{'}$ does not appear as the head entity in the relation r).

Firstly, we fuse the features of e_t. If e_t appears in the NF of e_h, then $\boldsymbol{Z}_1$ is used for fusion, otherwise $\boldsymbol{Z}_2$ is used for fusion.

$$\boldsymbol{Z}_1 = (1-\boldsymbol{W}_Z)\boldsymbol{M}_{m_{e_t}} + \boldsymbol{W}_Z \cdot \log \sum_{j=1}^{N_{e_t}} \exp \boldsymbol{M}_{m_{tj}} \tag{17}$$

$$\boldsymbol{Z}_2 = log \sum_{j=1}^{N_{e_t}} \exp \boldsymbol{M}_{m_{tj}} \tag{18}$$

$$\boldsymbol{P}_m = \left[\boldsymbol{Z}_{1/2}; \boldsymbol{e}_h; \boldsymbol{e}_h^{'}\right] \tag{19}$$

where $\boldsymbol{W}_Z \in \mathbb{R}^1$ is a learnable parameter.

3.3.2 Noun Fragment (NF) features

Further, we obtain features of the entity $e_h^{'}$ in the NF of e_h, including global and local features.

(i) Integrating global feature of the NF:

$$\boldsymbol{U}_{glo}^{pos_1,pos_2} = \frac{1}{pos_2 - pos_1 + 1} \sum_{k=pos_1}^{pos_2} \boldsymbol{M}_k \tag{20}$$

where pos_1 and pos_2 denote NF's start and end positions. $\boldsymbol{M}_k$ is the k-th feature obtained by Eq.(1). We fuse all words of the NF as the global feature.

(ii) Integrating local feature of the NF: Based on Eq.(9), the feature between the start word NF_{start} (or the end word NF_{end}) of the NF and the $e_h^{'}$ can be obtained:

$$\boldsymbol{V}^{'} = \boldsymbol{R}^{'}_{NF_{start},e_h^{'}} \cdot \boldsymbol{M} \tag{21}$$

$$\boldsymbol{V}^{''} = \boldsymbol{R}^{'}_{NF_{end},e_h^{'}} \cdot \boldsymbol{M} \tag{22}$$

We thus obtain the local feature of $e_h^{'}$ in the NF:

$$\boldsymbol{U}_{loc} = \boldsymbol{V}^{'}\boldsymbol{W}_{loc}\boldsymbol{V}^{''} + \boldsymbol{b}_{loc} \tag{23}$$

where $\boldsymbol{W}_{loc} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{b}_{loc} \in \mathbb{R}$ are learnable parameters.

(iii) The final relation feature of $(e_h^{'}, e_t)$ within the NF is as follows:

$$\boldsymbol{U}_{NF} = \left[\boldsymbol{U}_{loc}; \boldsymbol{U}_{glo}^{pos1,pos2}\right]\boldsymbol{W}_U + \boldsymbol{b}_U \tag{24}$$

where $\boldsymbol{W}_U \in \mathbb{R}^{2d\times d}$ and $\boldsymbol{b}_U \in \mathbb{R}$ are learnable parameters.

3.3.3 Mention reasoning

Finally, we fuse the mention features with the NF features to predict whether there is the relation r between the entity pair $(e_h^{'}, e_t)$.

$$\mathbb{P}_r^{'} = \mathrm{FFNN}([\boldsymbol{P}_m; \boldsymbol{U}_{NF}]) \tag{25}$$

3.4 Loss

As the model may have different confidence for different entity pairs, we apply the adaptive thresholding loss ATLOP (Zhou et al., 2021), which learns a dummy relation class TH during training that serves as a dynamic threshold for each relation class $r \in R$. For each (e_h, e_t), the loss forces the model to yield scores above TH for positive relation classes $R_P^{h,t}$ and scores below TH for negative relation classes $R_N^{h,t}$, formulated as below:

$$\begin{aligned}\mathcal{L}_r = -\sum_{h\neq t}\sum_{r\in R_P^{h,t}} \log\left(\frac{\exp\left(\mathbb{P}'_r\right)}{\sum_{r'\in R_P^{h,t}\cup\{\text{TH}\}}\exp\left(\mathbb{P}'_{r'}\right)}\right)\\ - \log\left(\frac{\exp\left(\mathbb{P}'_{\text{TH}}\right)}{\sum_{r'\in R_N^{h,t}\cup\{\text{TH}\}}\exp\left(\mathbb{P}'_{r'}\right)}\right)\end{aligned} \tag{26}$$

Our evidence module uses BCE loss:

$$\begin{aligned}\mathcal{L}_{evi} = -\sum_{h\neq t, NA\notin R_P^{h,t}}\sum_{i=1}^{N_s} y_n \cdot \boldsymbol{S}^i_{evi}\\ +(1-y_n)\cdot\log(1-\boldsymbol{S}^i_{evi})\end{aligned} \tag{27}$$

where $y_n \in \{0, 1\}$ is the evidence label.

4 Experiment and Analysis

4.1 Datasets and Implementation Settings

We assess our model on two most widely used datasets DocRED (Yao et al., 2019) and its revised version Re-DocRED (Tan et al., 2022b). We utilize the base versions of BERT (Devlin et al., 2019) and XLNET (Yang et al., 2019) as encoders. Details of *datasets* and *settings* are in **Appendix A.3**.

4.2 Evaluation Metrics and Baselines

We employ F1 and Ign F1 as main evaluation metrics following (Yao et al., 2019). Ign F1 is used to measure the F1 score while excluding relations shared between the training and dev/test sets. We also report Intra F1 and Inter F1 to evaluate abilities of intra-sentence and cross-sentence reasoning.

We compare with recent competitive models: (*i*) *Graph-based models*, including LSR, GAIN, CFER, HeterGSAN, MRN, DRN, SSAN, and SagDRE. (*ii*) *Transformer-based models*, including Coref, ATLOP, DocuNet, DocRE-II, Eider, KD-BERT, SAIS, SD-BERT, and DREEAM.

4.3 Main Results

As shown in **Table 1** and **Table 2**, our SRF-XLNET consistently outperforms all baselines on DocRED and Re-DocRED datasets and *achieves new state-of-the-art* (SOTA) *performance* in F1 and Ign F1. We also observe several interesting findings:

(*i*) Regardless of which encoder is used, our SRF outperforms all baselines on Re-DocRED, and *achieves greater gains compared with* the improvements on DocRED. Our SRF-XLNET obtains improvements of **2.01** F1 on the dev set of Re-DocRED, and relatively small 0.24 F1 on the dev set of DocRED. This suggests that our framework with secondary reasoning performs better on Re-DocRED that includes more completely annotated entity pairs involving reasoning.

(*ii*) We found that DRN-XLNET's Intra F1 is 0.38 higher than ATLOP, while ATLOP-XLNET's Inter F1 is 0.68 higher than DRN-XLNET. This may suggest that graph-based methods excel at within-sentence relations, while transformer-based models are better at cross-sentence relations. Despite *not incorporating a graph inference module*, our model still obtains competitive performance.

(*iii*) Our experiments also *reveal an interesting finding*: for the document-level RE task containing lots of long sentences, models that employ the original basic version of XLNET as the encoder generally outperform those that use the original basic version of BERT. We will conduct a more detailed experimental analysis in Section 4.8.

4.4 Ablation Study

To validate the effectiveness of each module in our model, we present ablation experiment results in **Table 3** and **Table 4**. We conduct four variants:

(*i*) Using a unidirectional attention method (denoted as **Unidirectional Attention**) by removing the tail-to-head attention from Eq.(4) to replace our bidirectional attention, resulting in a 0.43 and 0.69 decrease in F1 score on two datasets;

(*ii*) Using the mean of entity mentions directly (denoted as **Mean**) to replace bidirectional attention, resulting in a 0.38 and 0.61 decrease in F1 on two datasets. The above variants indicate that *our bidirectional attention is the optimal choice* among three ways for learning relation representation;

(*iii*) Removing the secondary reasoning module (denoted as **No Secondary Reasoning**), our model's F1 score decrease by 0.41 and 0.54 on two datasets, indicating that *new entity pair relations*

Model	Dev				Test	
	F1	Ign F1	Intra F1	Inter F1	F1	Ign F1
LSR-BERT (Nan et al., 2020)	59.00	52.43	65.26	52.05	59.05	56.97
GAIN-BERT (Zeng et al., 2020)	61.22	59.14	67.10	53.90	61.24	59.00
CFER-BERT (Dai et al., 2020)	61.41	59.23	-	-	61.28	59.16
HeterGSAN-BERT (Xu et al., 2021c)	60.18	58.13	-	-	59.45	57.12
MRN-BERT (Li et al., 2021)	62.01	60.02	-	-	62.06	60.24
DRN-BERT (Xu et al., 2021b)	61.39	59.33	-	-	61.37	59.15
SagDRE-BERT (Wei and Li, 2022)	62.11	60.32	-	-	62.32	60.11
Coref-BERT (Ye et al., 2020)	57.51	55.32	-	-	56.96	54.54
SSAN-BERT (Xu et al., 2021a)	59.19	57.03	-	-	58.16	55.84
ATLOP-BERT (Zhou et al., 2021)	61.09	59.22	-	-	61.30	59.31
DocuNet-BERT (Zhang et al., 2021)	61.83	59.86	-	-	61.86	59.93
DocRE-II (Zhang et al., 2022)	62.74	60.75	69.14	55.54	62.65	60.68
Eider-BERT (Xie et al., 2022)	62.48	60.51	68.47	55.21	62.47	60.42
SAIS-BERT (Xiao et al., 2022)	62.96	59.98	-	-	62.77	60.96
KD-BERT (Tan et al., 2022a)	62.03	60.08	-	-	62.08	60.04
SD-BERT (Zhang et al., 2023)	62.81	60.85	68.67	**56.09**	62.85	60.91
DREEAM-BERT (Ma et al., 2023)†	62.55	60.51	-	-	62.49	60.03
DRN-XLNET (Xu et al., 2021b)*	61.83	59.79	68.21	53.76	61.90	59.72
ATLOP-XLNET (Zhou et al., 2021)*	61.79	59.90	67.83	54.44	61.88	59.71
DocuNet-XLNET (Zhang et al., 2021)*	62.27	60.35	68.43	54.75	62.14	60.05
Eider-XLNET (Xie et al., 2022)*	62.80	60.57	-	-	62.81	60.67
SAIS-XLNET (Xiao et al., 2022)*	63.09	61.09	-	-	62.86	60.68
DREEAM-XLNET (Ma et al., 2023)†	61.94	60.07	-	-	61.68	59.64
SRF-BERT	62.50	60.46	-	-	62.11	59.84
SRF-XLNET	**63.33**	**61.33**	**69.22**	55.71	**63.07**	**60.98**

Table 1: Main results(%) on DocRED. Results with BERT are reported from their original papers. Results with $*$ are our reproduction using XLNET and † indicates our reproduction without their distantly-supervised data. Best results are in bold, and the second best are underlined.

Model	Dev		Test	
	F1	Ign F1	F1	Ign F1
ATLOP-BERT (Zhou et al., 2021)*	72.98	72.09	72.24	71.36
Eider-BERT (Xie et al., 2022)*	71.85	70.59	71.61	70.93
SAIS-BERT (Xiao et al., 2022)*	73.93	72.10	73.77	71.48
DREEAM-BERT (Ma et al., 2023)†	71.91	70.60	71.23	70.89
SRF-BERT	**74.66**	**73.76**	**74.06**	**73.16**
ATLOP-XLNET (Zhou et al., 2021)*	76.22	75.47	75.90	75.17
Eider-XLNET (Xie et al., 2022)*	75.77	74.52	75.47	74.80
SAIS-XLNET (Xiao et al., 2022)*	76.27	74.17	75.91	73.69
DREEAM-XLNET (Ma et al., 2023)†	75.76	75.00	75.06	74.29
SRF-XLNET	**78.28**	**77.25**	**76.33**	**75.90**

Table 2: Main results(%) on Re-DocRED. Considering that some models were not evaluated on the relatively new dataset Re-DocRED, we re-implemented several representative models using BERT and XLNET. Results with $*$ are our reproduction, and † are our reproduction without their distantly-supervised data. Best results using BERT and XLNET are in bold.

Model	F1
Bidirectional Attention (our SRF)	63.33
Unidirectional Attention	62.90
Mean	62.93
No Secondary Reasoning	62.92
No Evidence Extraction	62.81

Table 3: Ablation study on the dev set of DocRED.

Model	F1
Bidirectional Attention (our SRF)	78.28
Unidirectional Attention	77.59
Mean	77.67
No Secondary Reasoning	77.74
No Evidence Extraction	76.91

Table 4: Ablation study on the dev set of Re-DocRED.

are indeed predicted when we perform secondary reasoning on the first relation extraction results;

(*iv*) Removing the evidence extraction task (denoted as **No Evidence Extraction**), our model's performance experience the large drop, with a decrease of 0.52 and 1.37 in F1 on two datasets, indicating that although simple, *our evidence extraction method is very helpful* for relation extraction.

Moreover, we conduct the ablation analysis on *top-k layers* of multi-head attention at encoding layer mentioned in Section 3.1, and the results are detailed in **Appendix A.4**.

4.5 Generality of Secondary Reasoning

In addition to being effective in our model, the secondary reasoning module is also effective in other advanced models, as shown in **Table 5**. Here,

Model	Dev		Test	
	F1	Ign F1	F1	Ign F1
Eider-XLNET	62.80	60.57	62.81	60.67
+Secondary Reasoning	**62.95**	**60.73**	**62.92**	**60.79**
Eider-BERT	62.44	60.44	62.30	60.16
+Secondary Reasoning	**62.53**	**60.52**	**62.42**	**60.28**
SAIS-XLNET	63.09	61.09	62.86	60.68
+Secondary Reasoning	**63.20**	**61.18**	**62.97**	**60.77**
SAIS-BERT	62.96	59.98	62.77	60.96
+Secondary Reasoning	**63.05**	**60.07**	**62.83**	**61.04**
ATLOP-XLNET	61.79	59.90	61.88	59.71
+Secondary Reasoning	**61.93**	**60.03**	**61.99**	**59.82**
ATLOP-BERT	61.13	59.12	60.41	58.20
+Secondary Reasoning	**61.21**	**59.20**	**60.47**	**58.26**

Table 5: Experiments of generality when incorporating secondary reasoning into other models on DocRED.

Dataset	Associated	Not Associated	Total
Train	1100	955	2055
Dev	3380	2603	5983

Table 6: Association between entities within NFs in DocRED: ‘Total’ denotes the number of entities having relations with others. Entities sharing same tail entity are ‘Associated’; otherwise, they are ‘Not Associated’.

we select Eider, SAIS and ATLOP, and re-run them following their original settings, plus our secondary reasoning module. We have two main findings:

- Incorporating secondary reasoning into these models result in improved performance, regardless of the encoder being BERT or XLNET.
- Although adding secondary reasoning to the models can improve performance, the gains are limited. **This is because the Noun Fragments in the document are limited, resulting in a low upper limit for secondary reasoning**.

Moreover, we conduct some *Error Analysis and Case Studies* in **Appendix** A.5, the cases show that when incorporating our secondary reasoning, the existing models and ours make correct prediction.

To further explore whether there are indeed *associations* between entities in NFs (Noun Fragments), we conduct an analysis of all entities within NFs in **Table** 6. Our secondary reasoning is performed on NFs. It can be seen that over 53.5% of the ‘total’ entities within NFs are associated. This also *explains the necessity of exploring secondary reasoning*.

Model	Number of parameters	Evi F1
Eider	4.767M	51.07
SRF	0.001k	45.12

Table 7: Parameter comparison with the representative evidence extraction model Eider (Xie et al., 2022).

Model	Time	Memory Usage
ATLOP-BERT	133min	10484M
EIDER-BERT	83min	21218M
SAIS-BERT	196min	31952M
SRF-BERT	164min	24446M

Table 8: Training cost of our model on DocRED.

4.6 Analysis of Evidence Extraction

We perform an analysis of *parameter overhead* of the evidence extraction module as shown in **Table** 7. In the evidence extraction task, Eider used 4.7 million times more parameters than us, resulting in a 6% Evi F1 performance improvement over our model on DocRED dev set. But it should be noted that, *we can obtain evidence scores for all sentences in one calculation*, while Eider can only predict evidence score for one sentence at a time. Moreover, we incorporate our evidence extraction into other models, which consistently results in improved performance, as shown in **Appendix** A.6.

4.7 Training Cost of SRF

We conduct further analysis on the training cost of our model SRF (time and memory usage) compared to previous methods as shown in **Table** 8. Note that, the relation and evidence extraction modules are trained together. Throughout training, the joint training time for these two modules is 157min, with the majority of the training time coming from our proposed bidirectional mention interaction due to the large number of mentions. The training time for the secondary reasoning module is only 7min.

4.8 Exploratory Analysis using Different Encoders for Document-level RE

As stated in Section 4.3, our experiments *reveal an interesting and potentially useful finding*: for DocRE task containing lots of long sentences, models that employ the original basic version of XLNET as the encoder generally outperform those that use the original basic version of BERT. To further validate this finding, we conduct two experiments on the original DocRED and our constructed hard dataset with almost longer sentences. The detailed results in **Appendix** A.7 validate the above finding. While we haven’t tested all pre-trained models (including variants and larger versions of BERT or XLNet, as well as other models), we hope this finding may offer some insights for future document extraction methods regarding encoder selection.

5 Conclusion

We propose a novel secondary reasoning framework SRF for DocRE. In SRF, we propose a bidirectional mention fusion method and design a simple yet effective evidence extraction way by introducing only a learnable parameter to jointly perform relation prediction. On this basis, for the first time, we propose a novel secondary reasoning idea that is carefully designed to further explore results of the first relation prediction. Experiments demonstrate SRF achieves SOTA performance. Further analyses demonstrate that secondary reasoning is effective and general. In future we will apply our framework to more models and explore more encoders, and conduct more extensive experiments on a wider range of datasets.

Limitations

For the performance bound of secondary reasoning, as our experiments and analysis in Section 4.5, the secondary reasoning is performed on Noun Fragments, on one hand, *the number of Noun Fragments in the dataset is finite*, on the other hand, as the model's initial prediction performance improves, more entities within the Noun Fragments may be predicted during the first reasoning. Therefore, this may result in diminishing performance gains from the secondary reasoning. **For the impact of encoders on model performance**, the performance of our model (as well as some other models) is slightly impacted by the pre-trained encoder utilized. As demonstrated by our experiments, the original basic version of XLNET is more competitive than the original basic version of BERT when processing documents containing long sentences in the document-level relation extraction task. However, in practice, *it is challenging to draw a clear boundary* between long and short sentences, different real-world data may be affected by the pre-trained encoder used, including XLNET, BERT, other pre-trained models, and their larger versions or variations not tested in this paper.

Acknowledgments. The authors sincerely thank the anonymous reviewers for their valuable comments and suggestions, which improved the paper. The work is supported by the National Natural Science Foundation of China (62276057), and Sponsored by CAAI-MindSpore Open Fund, developed on OpenI Community.

References

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. 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)*, pages 4925–4936.

Damai Dai, Jing Ren, Shuang Zeng, Baobao Chang, and Zhifang Sui. 2020. Coarse-to-fine entity representations for document-level relation extraction. *arXiv preprint arXiv:2012.02507*.

Julien Delaunay, Thi Hong Hanh Tran, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Nicolas Sidere, and Antoine Doucet. 2023. A comprehensive survey of document-level relation extraction (2016-2022). *arXiv preprint arXiv:2309.16396*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Markus Eberts and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 3650–3660.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski Aapo Kyrola Andrew Tulloch, and Yangqing Jia Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2022. Does recommend-revise produce reliable annotations? an analysis on missing instances in docred. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6241–6252.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. 2021. Mrn: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *Learning*.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. Dreeam: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1963–1975.

Guoshun Nan, Zhijiang Guo, Ivan Sekulić, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1546–1557.

Xingyu Peng, Chong Zhang, and Ke Xu. 2022. Document-level relation extraction via subgraph reasoning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4331–4337.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Volume 1, Long Papers*, pages 1171–1182.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2895–2905.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of ACL*.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8472–8487.

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. Hin: Hierarchical inference network for document-level relation extraction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 197–209.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of NAACL-HLT*, pages 872–884.

Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-local neural networks for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3711–3721.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *arXiv preprint arXiv:1909.11898*.

Ying Wei and Qi Li. 2022. Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2000–2008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: system demonstrations (EMNLP)*, pages 38–45.

Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *Research in Computational Molecular Biology: 23rd Annual International Conference (RECOMB)*, pages 272–284.

Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2022. Sais: Supervising and augmenting intermediate steps for document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2395–2409.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Evidence-enhanced document-level relation extraction. In *Findings of the Association for Computational Linguistics (ACL)*, pages 257–268.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021a. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 14149–14157.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021b. Discriminative reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1653–1663.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021c. Document-level relation extraction with reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 14167–14175.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems (NeurIPS)*, 32.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 764–777.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1753–1762.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640.

Liang Zhang and Yidong Cheng. 2022. A densely connected criss-cross attention network for document-level relation extraction. *arXiv preprint arXiv:2203.13953*.

Liang Zhang, Jinsong Su, Yidong Chen, Zhongjian Miao, Min Zijun, Qingguo Hu, and Xiaodong Shi. 2022. Towards better document-level relation extraction via iterative inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8306–8317.

Liang Zhang, Jinsong Su, Zijun Min, Zhongjian Miao, Qingguo Hu, Biao Fu, Xiaodong Shi, and Yidong Chen. 2023. Exploring self-distillation based relational reasoning training for document-level relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pages 13967–13975.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3999–4006.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2205–2215.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020. Document-level relation extraction with dual-tier heterogeneous graph. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 1630–1641.

Chao Zhao, Daojian Zeng, Lu Xu, and Jianhua Dai. 2022. Document-level relation extraction with context guided mention integration and inter-pair reasoning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Huiwei Zhou, Yibin Xu, Weihong Yao, Zhe Liu, Chengkun Lang, and Haibin Jiang. 2020. Global context-enhanced graph convolutional networks for document-level relation extraction. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 5259–5270.

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 (AAAI)*, volume 35, pages 14612–14620.

A Appendix

A.1 Related Work

The existing DocRE methods may be roughly divided into two categories, including graph-based methods and non-graph-based methods (Delaunay et al., 2023).

Graph-based DocRE. The graph-based DocRE method is first proposed by (Quirk and Poon, 2017) and a series of subsequent methods are developed, including EoG (Christopoulou et al., 2019), GLRE (Wang et al., 2020), RENET (Wu et al., 2019), LSR (Nan et al., 2020), DHG (Zhang et al., 2020), GCGCN (Zhou et al., 2020), GAIN (Zeng et al., 2020), MRN (Li et al., 2021), HeterGSAN (Xu et al., 2021c), DRN (Xu et al., 2021b), SSAN (Xu et al., 2021a), SGR (Peng et al., 2022), and SagDRE (Wei and Li, 2022). These methods construct graphs with with heuristics or dependency information, and use entities or mentions as nodes of the graphs. On this basis, they encode the graphs using graph neural networks to obtain entity representations, and then predict relations by reasoning on the graphs.

Non-graph-based DocRE. Considering that the transformer (Vaswani et al., 2017) architecture can implicitly model long-range dependencies, lots of non-graph-based DocRE models begin to perform implicit reasoning using pre-trained models instead of document-level graphs, including Coref (Ye et al., 2020), Hin (Tang et al., 2020), JEREX (Eberts and Ulges, 2021), DocuNet (Zhang et al., 2021), DocRE-II (Zhang et al., 2022), Dense-CCNet (Zhang and Cheng, 2022), CGM2IR (Zhao et al., 2022).

Typically, Two_Step (Wang et al., 2019) provides two-step process for DocRE. The first step is to predict whether or not two entities have a relation, the second step is to predict the specific relation. The work also first proposes the new dataset for DocRE called DocRED. ATLOP (Zhou et al., 2021) is an adaptive thresholding and localized context pooling model for DocRE. ATLOP enriches the entity embedding with additional context relevant to the current entity pair, enabling a more comprehensive representation, and making a giant leap for DocRE. Eider (Xie et al., 2022) is an evidence-enhanced framework, which uses a bilinear network to extract evidence and fuses the extracted evidence in inference to enhance DocRE. SAIS (Xiao et al., 2022) identifies coreference resolution, named entity recognition, pooled evidence retrieval, and fine-grained evidence retrieval to compose four intermediate steps involved in the reasoning process.

Moreover, DREEAM (Ma et al., 2023) and KD (Tan et al., 2022a) use a large portion of distantly-supervised data in DocRED made by aligning Wikipedia articles with Wikidata. In addition, several work KD (Tan et al., 2022a) and SD (Zhang et al., 2023) explore distillation-based methods for DocRE.

A.2 Algorithm of Extracting Noun Fragments

Algorithm 1 gives part of the algorithm for extracting noun fragments.

A.3 Datasets and Implementation Settings

Dataset statistics. We evaluate our model using DocRED (Yao et al., 2019) and Re-DocRED (Tan et al., 2022b), two widely used large-scale, human-annotated datasets for document-level relation extraction derived from Wikipedia and Wikidata.

DocRED, which is the first dataset for document-level relation extraction, comprises 3,053 training documents, 1,000 development documents, and 1,000 test documents, encompassing 96 relation types, 132,275 entities, and 56,354 relational facts. Furthermore, over 40.7% of relational facts need to be extracted from multiple sentences, and 61.1% of relation instances necessitate multiple reasoning skills. The dataset also includes supporting sentences for each relation instance as part of its annotations.

Considering that some relations in DocRED are not marked (Huang et al., 2022; Tan et al., 2022b), Re-DocRED (Tan et al., 2022b) relabeled 4053 documents in DocRED, of which 3053 documents are used as the training set, 500 documents are used as the development set, and 500 documents are used as the test set.

Algorithm 1 Algorithm NF_extraction

```
procedure NF_EXTRACTION($S$)
    $NF \leftarrow \emptyset$
    $current\_fragment \leftarrow \emptyset$
    for all $s \in S$ do
        for all $w \in s$ do
            if $w$ is a noun or a punctuation or part of a noun phrase or one of ["in", "of"] then
                add $w$ to $current\_fragment$
            else if $current\_fragment$ contains at least three entities then
                add the current fragment from the first entity position to the last entity position to $NF$
                reset $current\_fragment$
            else
                reset $current\_fragment$
            end if
        end for
    end for
    return $NF$
end procedure
```

Details of Implementation Settings. Our model is built using PyTorch and the Transformers library from Huggingface (Wolf et al., 2020). We train our model using an NVIDIA RTX A6000 GPU with a memory size of 48G. We utilize BERT-base (Devlin et al., 2019), and XLNET-base (Yang et al., 2019) as our foundational encoders.

We optimize our model with AdamW, setting the encoder learning rate to 5e-5 for the first 45 epochs and 2e-6 thereafter, with other parameters set to 1e-5. For the first 45 epochs, we apply linear warm-up (Goyal et al., 2017) to the initial 6% of steps, followed by cosine annealing (Loshchilov and Hutter, 2016) for the remaining 255 epochs. The batch size (number of documents per batch) is set at 4, with the ratio between relation extraction loss and evidence extraction loss set at 0.01. The period of cosine annealing is 4 epochs. In the secondary reasoning module, the learning rate is adjusted to 2e-4, and other parameters are the same as above. In addition, the pre-training model of the secondary reasoning module directly adopts the fine-tuned pre-training model in the relation extraction module. We implement early stopping based on the F1 score on the development set, with a maximum of 300 epochs.

Moreover, we did not set a special way to select η. We argue that a too small η may make many irrelevant sentences become evidence sentences, while a too large η may lose many evidence sentences. Additionally, our SRF-XLNET model uses the concatenation strategy in Eq.(6) and SRF-BERT model uses the summation strategy in Eq.(7).

A.4 Experiments on Top-k Attention Layers

We investigate the impact of top-k layers of multi-head attention at the encoding layer mentioned in Section 3.1. The results are shown in Table 9.

The optimal result is achieved when k is set to 10 layers. We analyze this is because setting the number of layers too low would result in a loss of features, while setting it too high would incorporate low attention scores into the calculation, lowering the overall attention score.

Top-k	F1
Top-12 Layer Attention	63.20
Top-11 Layer Attention	63.24
Top-10 Layer Attention	**63.33**
Top-9 Layer Attention	63.12
Top-8 Layer Attention	62.89

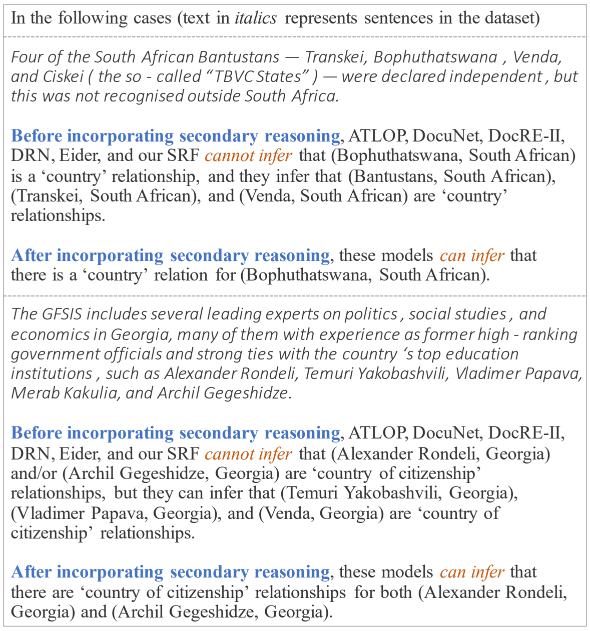
Table 9: Selection of k in the relation extraction module of Section 3.1 on DocRED dev set.

A.5 Case Studies of Secondary Reasoning

We conducted some error analysis and case studies as shown in **Figure 3**. In these cases, after incorporating our secondary reasoning, the existing models ATLOP, DocuNet, DRN, DocRE-II, Eider, and our SRF make the correct prediction. The cases also demonstrate that our secondary reasoning is effective and general.

A.6 Results of Incorporating Evidence Extraction into Other Models

In order to further demonstrate the generality and effectiveness of our evidence extraction, we conduct the experiment as shown in Table 10. The results show that our design of "evidence extraction" method consistently results in improved performance, when incorporating it into other models.



In the following cases (text in *italics* represents sentences in the dataset)

Four of the South African Bantustans — Transkei, Bophuthatswana , Venda, and Ciskei (the so - called "TBVC States") — were declared independent , but this was not recognised outside South Africa.

Before incorporating secondary reasoning, ATLOP, DocuNet, DocRE-II, DRN, Eider, and our SRF *cannot infer* that (Bophuthatswana, South African) is a 'country' relationship, and they infer that (Bantustans, South African), (Transkei, South African), and (Venda, South African) are 'country' relationships.

After incorporating secondary reasoning, these models *can infer* that there is a 'country' relation for (Bophuthatswana, South African).

The GFSIS includes several leading experts on politics , social studies , and economics in Georgia, many of them with experience as former high - ranking government officials and strong ties with the country 's top education institutions , such as Alexander Rondeli, Temuri Yakobashvili, Vladimer Papava, Merab Kakulia, and Archil Gegeshidze.

Before incorporating secondary reasoning, ATLOP, DocuNet, DocRE-II, DRN, Eider, and our SRF *cannot infer* that (Alexander Rondeli, Georgia) and/or (Archil Gegeshidze, Georgia) are 'country of citizenship' relationships, but they can infer that (Temuri Yakobashvili, Georgia), (Vladimer Papava, Georgia), and (Venda, Georgia) are 'country of citizenship' relationships.

After incorporating secondary reasoning, these models *can infer* that there are 'country of citizenship' relationships for both (Alexander Rondeli, Georgia) and (Archil Gegeshidze, Georgia).

Figure 3: Several case studies.

	Dev-F1	Dev-Ign F1
ATLOP-BERT	61.09	59.22
ATLOP-BERT+evi	61.58	59.53
DREEAM-BERT	62.55	60.51
DREEAM-BERT+evi	62.94	60.88
SAIS-BERT	62.96	59.98
SAIS-BERT+evi	63.27	61.15

Table 10: Experiments of the generality and effectiveness of incorporating the "Evidence Extraction" into other models on DocRED with BERT-base as encoder.

A.7 Results of Exploratory Experiments using Different Encoders for DocRE

Results on DocRED. Firstly, we segment all documents based on the length of the longest sentence contained in each document of DocRED dev set, and conduct experiments using our SRF and the existing Eider on different groups of documents. From **Figure 4**, we observe that documents consisting of short sentences achieve better performance with BERT, while XLNET performs better when documents contain longer sentences.

Results on Hard dataset. To further analyze and validate this finding, *we construct a challenging dataset*. We define sentences with 40 or more words as long sentences and extract documents with at least three long sentences to create the challenging dataset for experimentation. Through our experiments, we discover that our model effectively leverages performance potential of XLNET on documents with long sentences. As shown in **Figure**

5, our model exhibits the superiority in documents with long sentences, particularly outperforming ATLOP by a significant margin. We also find that models based on the XLNET encoder generally outperform those based on BERT on the challenging dataset.

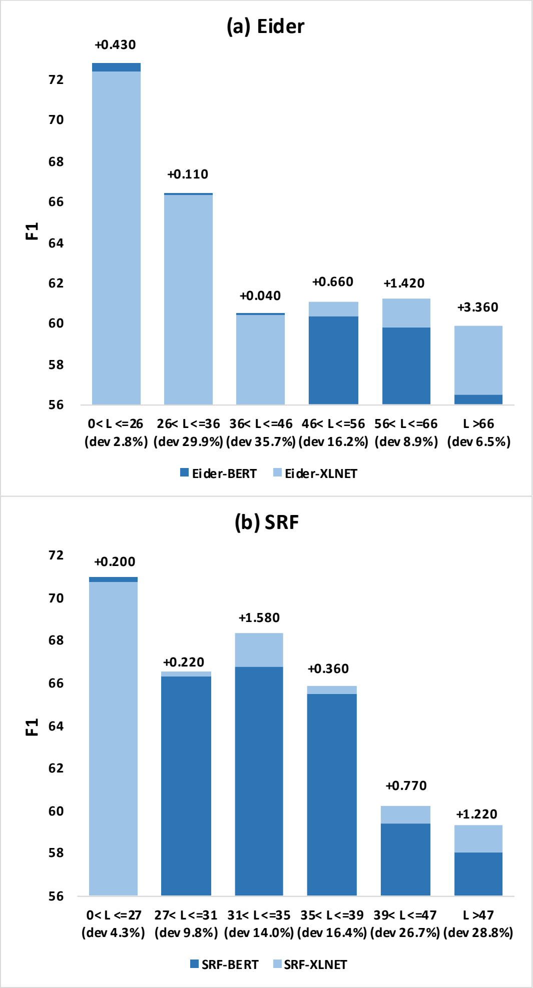


Figure 4: Performance of Eider and our SRF model on different groups of documents from DocRED dev set, using XLNET-base and BERT-base as encoders. "L" represents the length of the longest sentence in a document, and "*dev n%*" is the proportion of documents with the longest sentence length not exceeding L in 1000 documents of dev set. Moreover, considering that it is difficult to draw a clear boundary between long and short sentences, thus we set roughly consistent but different L values for models Eider and SRF to verify the performance.

Our analysis. All of the results on the DocRED dataset and our constructed hard dataset validate our finding: for DocRE task containing lots of long sentences, models that employ the original basic version of XLNET as the encoder generally outperform those that use the original basic version of BERT. We analyze and attribute this to the different encoding capabilities of XLNET and BERT for long sentences in documents. Our preliminary analysis may be because XLNET utilizes Transformer-XL techniques, which allow it to handle longer sequences by leveraging relative positional encoding and segment-level recurrence attention mechanism while retaining more historical information. Additionally, XLNET employs permutation language modeling, enabling it to randomly predict all words in a sequence rather than only predicting a subset of masked words like BERT. This helps avoid the inconsistency issue between pretraining and fine-tuning in BERT and allows for better capturing of global semantics in longer sentences. We will make more exploration and experiments to analyze the potential reasons in the future work. While we haven't tested all pre-trained models (including variants and larger versions of BERT or XLNet, as well as other models), we hope this finding may offer some insights for future document extraction methods regarding encoder selection.

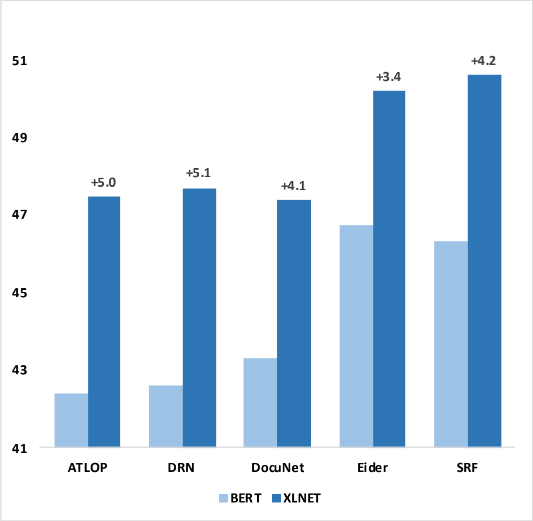


Figure 5: F1 Performance of our model SRF and several representative models on our constructed *hard dataset* using XLNET-base or BERT-base as the encoder.