Document-Level Relation Extraction with Sentences Importance Estimation and Focusing

Wang Xu[1], **Kehai Chen**[1], **Lili Mou**[2], **Tiejun Zhao**[1]
[1]Harbin Institute of Technology, China
[2]Dept. Computing Science, Alberta Machine Intelligence Institute (Amii)
University of Alberta, Canada
xuwang@hit-mtlab.net, {chenkehai,tjzhao}@hit.edu.cn, doublepower.mou@gmail.com

Abstract

Document-level relation extraction (DocRE) aims to determine the relation between two entities from a document of multiple sentences. Recent studies typically represent the entire document by sequence- or graph-based models to predict the relations of all entity pairs. However, we find that such a model is not robust and exhibits bizarre behaviors: it predicts correctly when an entire test document is fed as input, but errs when non-evidence sentences are removed. To this end, we propose a Sentence Importance Estimation and Focusing (SIEF) framework for DocRE, where we design a sentence importance score and a sentence focusing loss, encouraging DocRE models to focus on evidence sentences. Experimental results on two domains show that our SIEF not only improves overall performance, but also makes DocRE models more robust. Moreover, SIEF is a general framework, shown to be effective when combined with a variety of base DocRE models.[1]

1 Introduction

Document-level relation extraction (DocRE) aims to predict entity relations across multiple sentences. It plays a crucial role in a variety of knowledge-based applications, such as question answering (Sorokin and Gurevych, 2017) and large-scale knowledge graph construction (Baldini Soares et al., 2019). Different from sentence-level relation extraction (Zeng et al., 2014; Xiao and Liu, 2016; Song et al., 2019), the supporting evidence in the DocRE setting may involve multiple sentences scattering in the document. Thus, DocRE is more a realistic setting, attracting increasing attention in the field of information extraction.

Most recent DocRE studies use the entire document as a clue to predict the relations of all entity pairs without concerning where the evidence is located (Nan et al., 2020; Zeng et al., 2020; Xu et al., 2021a,b). However, one can identify the relation of a specific entity pair from a few sentences. Huang et al. (2021) show that irrelevant sentences in the document would hinder the performance of the model.

[1] *Rage Against the Machine* is an American rap metal band from Los Angeles, California. [2] Formed in 1991, the group consists of vocalist Zack de la Rocha, guitarist Tom Morello, bassist Tim Commerford and drummer *Brad Wilk*. [3] After a self-issued demo, the band signed with Epic Records and released its debut album *Rage Against the Machine* in 1992. …

Relation: MemberOf **Supporting Evidence:** {1,2}

Model Input	{1,2,3}	{1,2}
Ground Truth	MemberOf	MemberOf
GAIN Prediction	MemberOf	not MemberOf (undesired)

Figure 1: A DocRE model predicts correctly for an entire document, but errs when a non-evidence sentence is removed.

Moreover, we observe that a DocRE model, trained on the entire document, may err when non-evidence sentences are removed. In Figure 1, for example, we need to identify the relation "MemberOf" between the entities *Brad Wilk* and *Rage Against the Machine*. The evidence sentences are {1,2}, and humans can easily identify such a relation when reading sentences {1,2} only. However, the recent DocRE model GAIN (Zeng et al., 2020) identifies the relation "MemberOf" correctly from the entire document {1,2,3}, but predicts "not MemberOf" from sentences {1,2}. Intuitively, removing sentence {3} should not change the result, as this sentence does not provide information regarding whether "MemberOf" holds or not for the two entities. Such model behaviors are undesired, because it shows that the model is not robust and lacks interpretability.

To this end, we propose a novel **S**entence Importance **E**stimation and **F**ocusing (**SIEF**) framework to encourage the model to focus on evidence sentences for predicting the relation of

[1] The code is publicly available at `https://github.com/xwjim/SIEF`

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2920 - 2929
July 10-15, 2022 ©2022 Association for Computational Linguistics

an entity pair. Specifically, we first evaluate the importance of each sentence by the difference between the output probabilities of the document with and without this sentence. If the predicted probability of a relation does not change, or even increases, when a sentence is removed, it typically indicates that the sentence is *non-evidence*. Then, we propose an auxiliary loss to encourage the model to produce the same output distribution, when the entire document is fed as input and when a non-evidence sentence is removed. In this way, the model pays more attention to the evidence sentences for the classification. Our SIEF method is a general framework that can be combined with different underlying DocRE models.

We evaluated the generality and effectiveness of our approach on the large-scale DocRED dataset (Yao et al., 2019). Experimental results show that the proposed approach combines well with various recent DocRE models and significantly improves the performance. We further evaluated our approach on a dialogue relation extraction dataset, DialogRE (Yu et al., 2020); our SIEF yields consistent improvement, showing the generality of our approach in different domains.

2 Related Work

Relation extraction (RE) can be categorized by its granularity, such as sentence-level (Doddington et al., 2004; Xu et al., 2016; Wei et al., 2020) and document-level (Gupta et al., 2019; Zhu et al., 2019). Early work mainly focuses on sentence-level relation extraction. Pantel and Pennacchiotti (2006) propose a rule-based approach, and Mintz et al. (2009) manually design features for classifying relations. In the past several years, neural networks have become a prevailing approach for relation extraction (Xu et al., 2015; Song et al., 2019).

Document-level relation extraction (DocRE) is attracting increasing attention in the community, as it considers the interactions of entity mentions expressed in different sentences (Li et al., 2016; Yao et al., 2019). Compared with the sentence level, DocRE requires the model collecting and integrating inter-sentence information effectively. Recent efforts design sequence-based and graph-based models to address such a problem.

Sequence-based DocRE models encode a document by the sequence of words and/or sentences, for example, using the Transformer architecture (Devlin et al., 2019). Zhou et al. (2021) argue that the Transformer attentions are able to extract useful contextual features across sentences for DocRE, and they adopt an adaptive threshold for each entity pair. Zhang et al. (2021) model DocRE as a semantic segmentation task and predict an entity-level relation matrix to capture local and global information.

Graph-based DocRE models abstract a document by graphical structures. For example, a node can be a sentence, a mention, and/or an entity; their co-occurrence is modeled by an edge. Then graph neural networks are applied to aggregate inter-sentence information (Quirk and Poon, 2017; Christopoulou et al., 2019; Zeng et al., 2020). Zeng et al. (2020) construct double graphs, applying graph neural networks to mention–document graphs and performing path reasoning over entity graphs. Xu et al. (2021a) explicitly incorporate logical reasoning, common-sense reasoning, and coreference reasoning into DocRE, based on both sequence and graph features.

Different from previous work, our paper proposes SIEF as a general framework that can be combined with various sequence-based and graph-based DocRE models. In our approach, we propose a sentence importance score and a sentence focusing loss to encourage the model to focus on evidence sentences, improving the robustness and the overall performance of DocRE models.

3 Problem Definition

In this section, we present the formulation of document relation extraction (DocRE). Consider an unstructured document comprising N sentences, $\mathbb{D} = \{\mathrm{s}_1, \mathrm{s}_2, \cdots, \mathrm{s}_N\}$, where each sentence s_n is a sequence words. In a DocRE dataset, the document $\mathbb{D}$ is typically annotated with entity mentions, each mention (e.g., U.S. and USA) labeled by its conceptual entity e and its entity type (e.g., location).

A DocRE model $\boldsymbol{F}$ is usually formulated as multi-label classification (Yao et al., 2019). $\boldsymbol{F}_j$ predicts whether the jth relation holds for the ith marked entity pair in a document, given by

$$P_{ij} = \boldsymbol{F}_j(\mathbb{D}, e_{i_h}, e_{i_t}) = \Pr[r_{ij} = 1 | \mathbb{D}, e_{i_h}, e_{i_t}] \tag{1}$$

where e_{i_h} is the head entity and e_{i_t} is the tail entity; $r_{ij} \in \{0, 1\}$ is the groundtruth label regarding entity pair i and relation j.

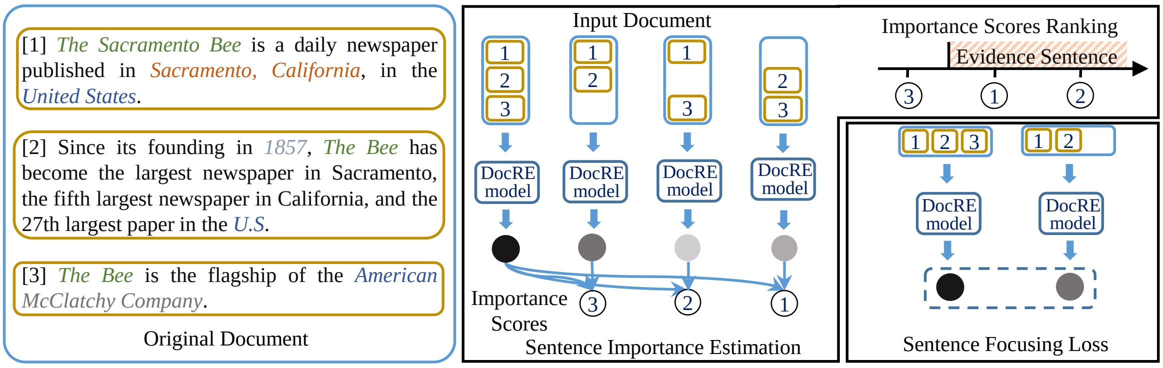

[1] The Sacramento Bee is a daily newspaper published in Sacramento, California, in the United States.
[2] Since its founding in 1857, The Bee has become the largest newspaper in Sacramento, the fifth largest newspaper in California, and the 27th largest paper in the U.S.
[3] The Bee is the flagship of the American McClatchy Company.
Original Document
Input Document
1
2
3
DocRE model
Importance Scores
Sentence Importance Estimation
Importance Scores Ranking
Evidence Sentence
DocRE model
Sentence Focusing Loss

Figure 2: We estimate the sentence importance (for a specific entity pair and relation) by the difference of the classification probabilities with and without the sentence. Then, we encourage the DocRE model to predict the same probability when the entire document is fed as input and when a non-evidence sentence is removed.

To train the model, the binary cross-entropy loss is used as the objective for parameter estimation:

$$\mathcal{L}_{\text{rel}} = -\sum_{\mathbb{D}\in\mathcal{C}}\sum_{i_h\neq i_t}\sum_{j\in\mathcal{R}}\{r_{ij}\log P_{ij} + (1-r_{ij})\log(1-P_{ij})\} \tag{2}$$

where $\mathcal{C}$ denotes the entire corpus and $\mathcal{R}$ denotes the set of relation types.

During inference, we obtain the relation(s) of a given entity pair by thresholding the predicted probabilities, following most previous work (Yao et al., 2019; Zhou et al., 2021).

4 Methodology

In this section, we will describe our approach in detail. The overview of our framework is shown in Figure 2. First, we describe the estimation of sentence importance in Section 4.1. Sentences with low importance scores are treated as non-evidence. Then, Sections 4.2 and 4.3 present our approach that encourages the model to produce the same output distribution, when the entire document is fed as input and when non-evidence sentences are removed. Section 4.4 further presents the architectures of DocRE models.

4.1 Sentence Importance Estimation

We estimate the importance of each sentence for a specific entity pair. Low-scored sentences will be treated as non-evidence, and in principle, can be removed without changing DocRE predictions.

We propose a sentence importance score based on the DocRE predictions with and without the sentence in question. Our observation is that the relation extraction task is usually monotonic to evidence, i.e., (non-strictly) more relations will be predicted with more sentences. If we remove a sentence and the predicted probability of a relation decreases, then the sentence is likely to be the evidence. If the predicted probability does not change, then the sentence is likely to be non-evidence. Moreover, the predicted probability may sometimes increase when a sentence is removed, in which case the DocRE model is not robust, as this violates monotonicity.

Formally, we consider removing one sentence at a time, and the document with the nth sentence removed is denoted by $\hat{\mathbb{D}}^{(-n)} = \{\mathbf{s}_1, \cdots, \mathbf{s}_{n-1}, \mathbf{s}_{n+1}, \cdots, \mathbf{s}_N\}$. For a DocRE model $\boldsymbol{F}$, we obtain the classification probabilities $P_{ij} = \boldsymbol{F}_j(\mathbb{D}, e_{i_h}, e_{i_t})$ based on the original document, and $\hat{P}_{ij}^{(-n)} = \boldsymbol{F}_j(\hat{\mathbb{D}}^{(-n)}, e_{i_h}, e_{i_t})$ with sentence n removed.

We propose the importance score as

$$g_{ij}^{(-n)} = P_{ij}\log\frac{P_{ij}}{\hat{P}_{ij}^{(-n)}} \tag{3}$$

The formula appears similar to Kullback–Leibler (KL) divergence. However, we only take one term in the KL summation, because the KL divergence, albeit asymmetric in its two arguments, cannot model the increase or decrease of $\hat{P}_{ij}^{(-n)}$, whereas our $g_{ij}^{(-n)}$ is monotonically decreasing with $\hat{P}_{ij}^{(-n)}$. Compared with a naive difference or ratio between P_{ij} and $\hat{P}_{ij}^{(-n)}$, we find that our KL-like score is more robust in the scale of P_{ij} when determining non-evidence sentences.

We treat a sentence n as *non-evidence* if $g_{ij}^{(-n)} < \beta$ for a thresholding hyperparameter β. The resulting set of non-evidence sentences is denoted by $\mathbb{K}_{ij}$ for the an entity pair (e_{i_h}, e_{i_t}) and relation j.

4.2 Sentence Focusing Loss

We propose a sentence focusing loss to encourage the model to produce the same output distribution when the entire document is fed as input and when non-evidence sentences are removed.

Ideally, the predicted probability should remain the same if we remove any combination of the sentences in $\mathbb{K}_{ij}$. Therefore, we penalize the extent to which the predicted probability is changed.

We propose the sentence focusing loss as:

$$\mathcal{L}_{\text{sf}} = -\sum_{\mathbb{D}\in\mathcal{C}}\sum_{i_h\neq i_t}\sum_{j\in\mathcal{R}}\sum_{\mathbb{J}_{ij}\subseteq\mathbb{K}_{ij}} \{P_{ij}\log(\hat{P}_{ij}^{(-\mathbb{J}_{ij})}) \\ +(1-P_{ij})\log(1-\hat{P}_{ij}^{(-\mathbb{J}_{i})})\} \tag{4}$$

where $\mathbb{J}_{ij}$ is a subset of $\mathbb{K}_{ij}$ and $\hat{P}_{ij}^{(-\mathbb{J}_{ij})} = \boldsymbol{F}_j(\mathbb{D}\backslash\mathbb{J}_{ij}, e_{i_h}, e_{i_t})$ is the predicted probability with $\mathbb{J}_{ij}$ removed from $\mathbb{D}$, and the total loss is $\mathcal{L} = (\mathcal{L}_{\text{rel}} + \mathcal{L}_{\text{sf}})/2$.

Essentially, our sentence focusing loss ensures P_{ij} is close to $\hat{P}_{ij}^{(-\mathbb{J}_{ij})}$, which intuitively makes sense because non-evidence sentences should not affect the prediction. Our approach can also be thought of as a way of data augmentation. However, compared with one-hot groundtruth labels, our sentence focusing loss works with soft labels P_{ij} and $\hat{P}_{ij}^{(-\mathbb{J}_{ij})}$, which are believed to contain more information (Hinton et al., 2015), and our gradient propagates to both P_{ij} and $\hat{P}_{ij}^{(-\mathbb{J}_{ij})}$ for training.

The calculation of Eqn. (4) is time- and resource-consuming, because the number of the subsets $\mathbb{J}_{ij}$ grows combinatorially with the number of non-evidence sentences. Moreover it should be calculated repeatedly once the parameter of the model is updated. To this end, we propose a simplified training strategy to approximate Eqn. (4) in the next subsection.

4.3 Training Strategy

We propose a strategy to simplify the calculation and the training procedure. Concretely, we only remove one non-evidence sentence in $\mathbb{K}_{ij}$ at a time instead of a subset of $\mathbb{J}_{ij} \subseteq \mathbb{K}_{ij}$, and we aggregate the effect of different non-evidence sentences by:

$$\mathcal{L}_{\text{sf}} = -\sum_{\mathbb{D}\in\mathcal{C}}\sum_{n=1}^{N}\sum_{i_h\neq i_t}\sum_{j\in\mathcal{R}}\mathbb{I}(g_{ij}^{(-n)} < \beta) \\ \{P_{ij}\log(\hat{P}_{ij}^{(-n)}) + (1-P_{ij})\log(1-\hat{P}_{ij}^{(-n)})\} \tag{5}$$

where $\mathbb{I}$ is the indicator function. Essentially, we linearly approximate the combination of multiple non-evidence sentences in (4) by an outer summation. In this way, the number of terms does not grow combinatorially, but linearly w.r.t. N.

In implementation, we further simply the summation over n by Monte Carlo sampling of a randomly selected sentence n in each gradient update. The loss is reformulated as follows:

$$\mathcal{L}_{\text{sf}} = -\sum_{\mathbb{D}\in\mathcal{C}}\sum_{i_h\neq i_t}\sum_{j\in\mathcal{R}}\mathbb{I}(g_{ij}^{(-n)} < \beta) \\ \{P_{ij}\log(\hat{P}_{ij}^{(-n)}) + (1-P_{ij})\log(1-\hat{P}_{ij}^{(-n)})\} \tag{6}$$

As seen, we need to forward the base models twice in each update, with and without the sentence n. Huang et al. (2021) propose a similar idea but train different entity pairs in a document based on different sets of sentences; all sentence are processed repeatedly among entity pairs in a document. Their approach is much slower than ours.

To sum up, the proposed SIEF framework identifies non-evidence sentences and penalizes the difference of predicted probabilities when a non-evidence sentence is removed. Our approach is a generic framework and can be adapted to various DocRE model easily, without introducing extra parameters into the model.

4.4 DocRE Model Architectures

Our SIEF can be applied to various base DocRE models. To evaluate its generality, we consider the following recent models.

BiLSTM (Yao et al., 2019)[2]. A bi-directional long short term memory (BiLSTM) encodes the document, and an entity is representated by BiLSTM's hidden states, averaged over entity mentions. The head and tail entity representations are fed to a multi-layer perceptron (MLP) for relation extraction.

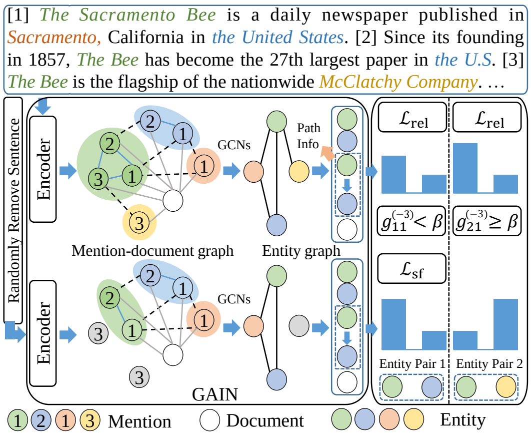
BERT$_{\text{base}}$ (Devlin et al., 2019)[3]. A pre-trained language model is used for document encoding.

HeterGSAN (Xu et al., 2021b)[4]. HeterGSAN is a recent graph-based DocRED model, which constructs a heterogeneous graph of sentence, mention, and entity nodes; it uses graph neural networks for relation extraction.

[2] https://github.com/thunlp/DocRED

[3] https://github.com/DreamInvoker/GAIN

[4] https://github.com/xwjim/DocRE-Rec



[1] *The Sacramento Bee* is a daily newspaper published in *Sacramento,* California in *the United States*. [2] Since its founding in 1857, *The Bee* has become the 27th largest paper in *the U.S*. [3] *The Bee* is the flagship of the nationwide *McClatchy Company*. …

Figure 3: The model architecture of GAIN with SIEF. A sentence is randomly removed from the document. The corresponding nodes and edges are removed from the mention–document graph and the entity graph.

GAIN (Zeng et al., 2020)[3]. GAIN constructs two graphs: mention–document graphs and entity graphs, and performs graph and path reasoning over the two graphs separately. When combining our SIEF with GAIN, we achieve the best performance among all the base models with SIEF on DocRED. Thus, we will explain this model in more detail.

Essentially, a node in the mention–document graph is either a mention or a document. The mentions are connected to its document, and two mentions are connected if they co-occur in one sentence. In the entity graph, two entities are connected if they are mentioned in one sentence. To classify the relation, GNN is applied to the mention–document graph, enhanced with path information in the entity graph, shown in Figure 3.

When combining SIEF with GAIN, we randomly remove one sentence from the document. The corresponding nodes and edges are removed in the GAIN's graphs. Then we obtain the output probabilities with and without the sentence, P_{ij} and $\hat{P}_{ij}^{(-n)}$, separately. If the sentence important score $g_{ij}^{(-n)}$ in Eqn. (3) is below a threshold β, the sentence is treated as non-evidence for the entity pair (e_{i_h}, e_{i_t}) and relation j. We apply the sentence focusing loss Eqn. (4) to improve the robustness.

For prediction, we apply the trained DocRE model to the entire document, because with our approach the model is already robust when non-evidence sentences are presented. Empirical results will show that our SIEF consistently improves the performance of base DocRE models.

5 Experiments

5.1 Setup

Datasets. DocRED is a large-scale human-annotated dataset for document-level relation extraction (Yao et al., 2019). The dataset is constructed from Wikipedia and Wikidata, containing 3053 documents for training, 1000 for development, and 1000 for test. In total, it has 132,375 entities and 56,354 relational facts in 96 relation types. More than 40% of the relational facts require reasoning over multiple sentences. The standard evaluation metrics are F1 and Ign F1 (Yao et al., 2019; Zeng et al., 2020), where Ign F1 refers to the F1 score excluding the relational facts in the training set.

We also evaluated our approach on DialogRE (V2, Yu et al., 2020), which contains 36 relation types, 17 of which are interpersonal. We followed the standard split with 1073 training dialogues, 358 validation, and 357 test. Following Yu et al. (2020), we report macro F1 scores in both the standard and conversational settings; the latter is denoted by F1$_c$.

Competing Methods. We experimented our SIEF on a number of base models, namely, BiLSTM, BERT$_{\text{base}}$, HeterGSAN, and GAIN (Section 4.4). These base models are all considered for comparison.

For DocRED, we consider additional competing methods: **Two Phase** (Wang et al., 2019), which first predicts whether the entity pair has a relation and then predicts the relation type; **LSR** (Nan et al., 2020), which constructs the graph by inducing a latent document-level graph; **Reconstructor** (Xu et al., 2021b), which encourages the model to reconstruct a reasoning path during training; **DRN** (Xu et al., 2021a), which considers different reasoning skills explicitly and uses graph representation and context representation to model the reasoning skills; **ATLOP** (Zhou et al., 2021), which aggregates contextual information by the Transformer attentions and adopts an adaptive threshold for different entity pairs; and **DocuNet** (Zhang et al., 2021), which models DocRE as a semantic segmentation task.

For DialogRE, we followed Yu et al. (2020) and considered **BERT** and **BERT**$_{\text{s}}$ for comparison,[5] where **BERT**$_{\text{s}}$ prevents a model from overfitting by replacing of the interpersonal augment with a special token.

[5] https://github.com/nlpdata/dialogre

Model	Dev		Test	
	Ign F1	F1	Ign F1	F1
DocRE Systems with GloVe				
LSR (Nan et al., 2020)	48.82	55.17	52.15	54.18
Reconstructor (Xu et al., 2021b)	54.25	55.70	53.25	55.13
DRN (Xu et al., 2021a)	54.61	56.49	54.35	56.33
BiLSTM (Yao et al., 2019)	48.87	50.94	48.78	51.06
+SIEF	52.08	54.20	51.03	53.22
HeterGSAN (Xu et al., 2021b)	52.17	54.40	52.07	53.52
+SIEF	54.49	56.30	53.94	55.85
GAIN (Zeng et al., 2020)	53.05	55.29	52.66	55.08
+SIEF	55.07	56.96	54.72	56.75
DocRE Systems with BERT$_{base}$				
Two-Phase (Wang et al., 2019)	-	54.42	-	53.92
LSR (Nan et al., 2020)	52.43	59.00	56.97	59.05
Reconstructor (Xu et al., 2021b)	58.13	60.18	57.12	59.45
DRN (Xu et al., 2021a)	59.33	61.39	59.15	61.37
ATLOP (Zhou et al., 2021)	59.22	61.09	59.31	61.30
DocuNet (Zhang et al., 2021)	59.86	61.83	**59.93**	61.86
BERT$_{\text{base}}$ (Ye et al., 2020)	54.63	56.77	53.93	56.27
+SIEF	57.13	59.11	57.87	58.93
HeterGSAN (Xu et al., 2021b)	57.00	59.13	56.21	58.54
+SIEF	57.99	60.04	57.93	60.02
GAIN (Zeng et al., 2020)	59.14	61.22	59.00	61.24
+SIEF	**59.82**	**62.24**	59.87	**62.29**

Table 1: Results on the development and test sets of the DocRE dataset. Bold indicates the best performance.

Model	Dev		Test	
	F1	F1$_c$	F1	F1$_c$
BERT (Yu et al., 2020)	60.6	55.4	58.5	53.2
+SIEF	61.4	57.6	59.9	56.1
BERT$_{\text{s}}$ (Yu et al., 2020)	63.0	57.3	61.2	55.4
+SIEF	**64.3**	**60.6**	**61.8**	**58.4**

Table 2: Results on DialogRE.

Implementation Details. We use the repositories[2,3,4,5] of base models to implement our approach. We mostly followed the standard hyperparameters used in the base models. Our SIEF has one hyperparameter β in Eqn. (5). It was set to 0.8, and Section 5.2 presents the effect of tuning β.

5.2 Results and Analyses

Main results. Table 1 presents the detailed results on the development and test sets of the DocRED dataset. We first compare DocRE systems with GloVe embeddings (Yao et al., 2019). We see that the proposed SIEF method significantly improves the performance of all base models, including the sequence model (i.e., BiLSTM) and graph models (i.e., HeterGSAN and GAIN); the average improvement is 2.05 points in terms of test F1. This shows that SIEF is compatible with both sequence and graph models, indicating the generality and effectiveness of the proposed method.

For the DocRE system with BERT$_{\text{base}}$, SIEF also consistently improves the base models, showing that SIEF is complementary to the modern BERT architecture. Especially, combining SIEF and GAIN (Zeng et al., 2020) with BERT$_{\text{base}}$ encoding yields state-of-the-art performance in terms of F1.

Model	Intra-F1	Inter-F1
BiLSTM	57.05	43.49
+SIEF	60.56 $_{(\Delta=+3.51)}$	45.96$_{(\Delta=+2.47)}$
HeterGSAN	61.79	47.06
+SIEF	63.01 $_{(\Delta=+1.22)}$	48.11$_{(\Delta=+1.05)}$
GAIN	61.67	48.77
+SIEF	**63.21** $_{(\Delta=+1.54)}$	**48.98**$_{(\Delta=+0.21)}$

Table 3: Results of Intra-F1 results and Infer-F1 on development set of DocRED. The difference is compared between SIEF and the respective base model.

We further conducted experiments on the DialogRE dataset, and compare our approach with the BERT baselines in Yu et al. (2020). As seen, the results are consistent with the improvement on DocRED, as our SIEF largely improves F1 and F1$_c$ for both base models. This further confirms the generality of our approach in different domains.

In the rest of this section, we present in-depth analyses to better understand our model with DocRED as the testbed. All base models use GloVe embeddings as opposed to BERT due to efficiency concerns.

Intra- and Inter-Sentence Performance. We breakdown the relation classification performance into intra-sentence reasoning and inter-sentence reasoning. Ideally, if only one sentence is needed to determine the relation of an entity pair, then it belongs to the intra-sentence category; if two or more sentences are needed, then it belongs to the inter-sentence category. We follow Nan et al. (2020) and approximate it by checking whether two entities are mentioned in one sentence.

The results are shown in Table 3. SIEF again consistently improves base models in terms of both Intra-F1 and Inter-F1. However, the improvement on Intra-F1 is larger than that on Inter-F1. This is because our SIEF encourages the model to focus on evidence by removing one sentence at a time, but does not explicitly model sentence relations. Based on this analysis, we plan to extend the SIEF framework with multi-sentence DocRE reasoning in our future work.

Performance of predicting evidence sentences. In our paper, we propose a sentence importance score to measure how much a sentence contributes to the classification without using additional

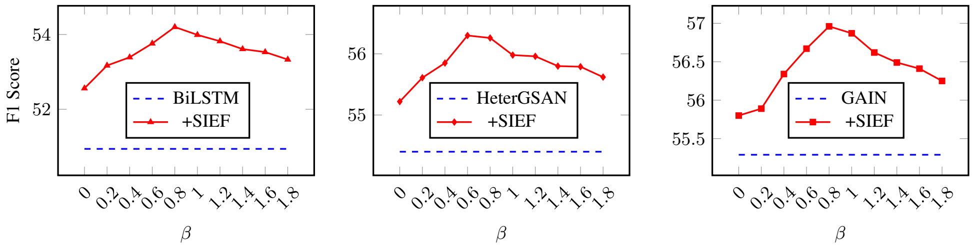


Figure 4: Performances of the classification (in F1 scores) on the development set of different hyperparameter β in Eqn. (6) during the training.

Model	Precision	Recall	F1
BiLSTM	60.14	68.41	64.01
+SIEF	65.00	67.99	66.46$_{(\Delta=+2.45)}$
HeterGSAN	65.40	70.95	68.06
+SIEF	71.40	70.21	70.80$_{(\Delta=+2.74)}$
GAIN	65.28	71.17	68.10
+SIEF	**71.94**	**71.60**	**71.77**$_{(\Delta=+3.67)}$

Table 4: Results of the evidence prediction on the development set of DocRED.

annotation. We evaluate such performance in Table 4 by Precision, Recall, and F1 scores against manually annotated evidence sentences that are provided in the dataset. In this analysis, we do not perform relation prediction, but concern about entity pairs knowingly having certain relations. Specifically, for entity pair (e_{i_h}, e_{i_t}) with relation j, we calculate the importance score $g_{ij}^{(-n)}$ for each sentence and cut off evidence/non-evidence sentences with a threshold based on the development F1 score.

As seen, all base models achieve above 60% F1, suggesting that the proposed importance score is indeed indicative for predicting evidence and non-evidence sentences.

With the proposed SIEF framework, the performance improves for all metrics, with an average improvement of 2.95 F1 points across three base models. This further verifies that our SIEF framework not only improves relation extraction performance, but also is able to better detect evidence and non-evidence sentences, which is important for the interpretability of machine learning models.

Robustness of DocRE models. We further investigate the robustness of DocRE models by showing the difference between the predicted distributions with and without non-evidence sentences. We show in Figure 5 the scatter plots of the probability P based on the entire document and the probability $\hat{P}_{ij}^{(-n)}$ with a random non-evidence sentence removed.

As shown in the figure, the points of the base models (left magenta plots) scatters over a wider range, whereas our SIEF training (right cyan plots) makes them more concentrated on the diagonal, indicating that the prediction P_{ij} on the entire document is mostly the same as $\hat{P}_{ij}^{(-n)}$ with a non-evidence removed. This shows the robustness of SIEF-trained models, as they are less sensitive to non-evidences sentences for DocRE.

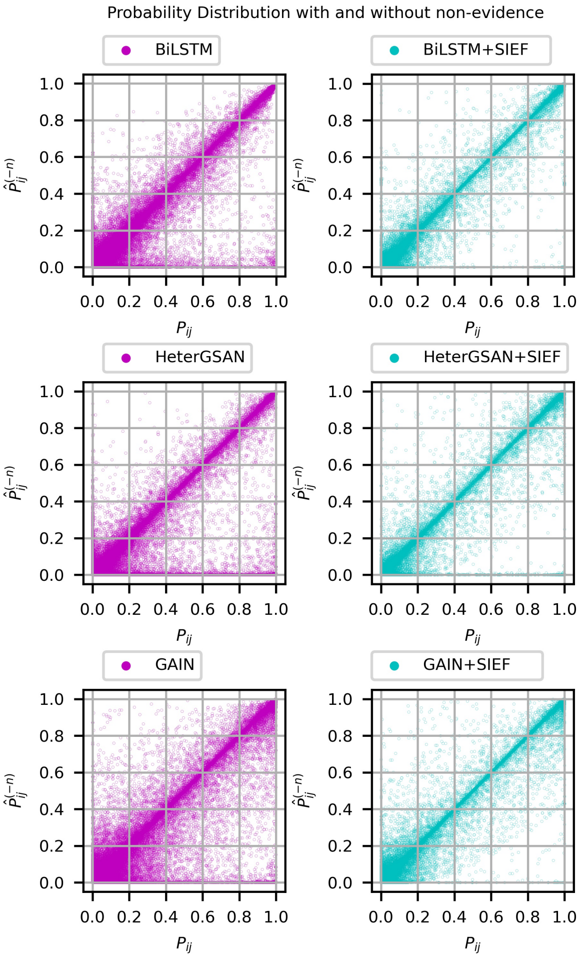


Figure 5: Robustness of DocRE models.

Analysis on hyperparameter β. Our SIEF framework has one hyperparameter β that controls how strict we treat a sentence as evidence or non-evidence (Section 4.3). We analyze the effect of β in Figure 4.

Method	BiLSTM		HeterGSAN		GAIN	
	Ign F1	F1	Ign F1	F1	Ign F1	F1
Base	48.87	50.94	52.17	54.40	53.05	55.29
+SIEF	**52.08**	**54.20**	**54.49**	**56.30**	**55.07**	**56.96**
+Rand	50.63	52.63	52.75	54.70	53.41	55.63
+NoMention	51.56	53.79	54.07	55.95	54.66	56.52

Table 5: Results of our approach and other heuristics.

Method	BiLSTM		HeterGSAN		GAIN	
	Ign F1	F1	Ign F1	F1	Ign F1	F1
Base	48.87	50.94	52.17	54.40	53.05	55.29
+SIEF	**52.08**	**54.20**	**54.49**	**56.30**	**55.07**	**56.96**
+GTruth	50.36	52.56	52.65	54.69	53.75	55.87

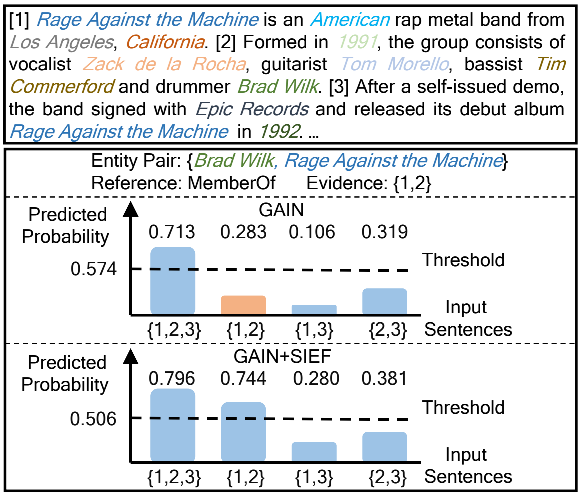
Table 6: Comparing our sentence focusing loss with learning from groundtruth labels (denoted by GTruth).

As seen, our SIEF approach consistently benefits the base models with a large range of β values. Intuitively, if β is too small, very few sentences will be treated as non-evidence and our sentence focusing loss is less effective; if β is too large, it has a high false positive rate of non-evidence sentences. Empirically, a moderate β around (0.6–0.8) yields the highest performance. From the plots, we also see that our hyperparameter β is insensitive to the base models, justifying our design of Eqn. (3).

Sentence importance score VS other heuristics. To investigate the effectiveness of our sentence importance score in Eqn. (3), we compare it with several alternative heuristics: 1) We randomly select half of the sentences as the non-evidence set, denoted by **Rand**; and 2) We consider the non-evidence set as the sentences without entity mentions, denoted by **NoMention**.

The results of the performance in terms of F1 and Ign F1 on the development set are shown in Table 5. As seen, the simple heuristic Rand outperforms the base model, as Rand can be thought of as noisy data augmentation. The NoMention heuristic outperforms Rand, as sentences without entity mentions are more likely to be non-evidence. Moroever, SIEF is superior to both Rand and NoMention, showing that our sentence importance scores is a more effective indicator of evidence and non-evidence sentences.

Our sentence focusing loss VS learning from groundtruth. We encourage the DocRE models to generate consistent output probabilities with and without non-evidence (Section 4.2) by a cross-entropy loss between two soft distributions P_{ij} and $\hat{P}_{ij}^{(-n)}$. To investigate the effect of such a sentence focusing loss, we compare it with an alternative choice: we learn $\hat{P}_{ij}^{(-n)}$ directly from the groundtruth label r_{ij}.


[1] *Rage Against the Machine* is an *American* rap metal band from *Los Angeles*, *California*. [2] Formed in *1991*, the group consists of vocalist *Zack de la Rocha*, guitarist *Tom Morello*, bassist *Tim Commerford* and drummer *Brad Wilk*. [3] After a self-issued demo, the band signed with *Epic Records* and released its debut album *Rage Against the Machine* in *1992*. ...

Entity Pair: {*Brad Wilk*, *Rage Against the Machine*}
Reference: MemberOf Evidence: {1,2}

Figure 6: Case Study.

Table 6 shows the results on the development set in terms of F1 and Ign F1. As seen, both methods can improve the performance of the base models. This confirms that removing non-evidence sentences can serve as a way of data augmentation, boosting the performance of DocRE models. Moreover, we observe that our sentence focusing loss is better than learning from the groundtruth labels, showing that the soft predictions provide more information than one-hot labels, consistent with knowledge distillation literature (Hinton et al., 2015).

Case Study. Figure 6 shows a case study of GAIN and GAIN+SIEF models. For the entity pair (*Brad Wilk*, *Rage Against the Machine*), both GAIN and GAIN+SIEF predicts the relation “MemberOf”, which is consistent with the reference. We see that Sentence 3 is non-evidence, and in principle, it should not affect DocRE prediction in this case. However, the base GAIN model makes a wrong prediction “not MemberOf”, as the predicted probability is below the threshold, which is determined by validation based on predicted binary probabilities of all relations. By contrast, our SIEF model is able to make correct predictions when different non-evidence sentences are removed, demonstrating its robustness.

6 Conclusion

In this paper, we propose a novel Sentence Information Estimation and Focusing (SIEF) approach to document relation extraction (DocRE).

We design a sentence importance score and a sentence focusing loss to encourage the model to focus on evidence sentences. The proposed SIEF is a general framework, and can be combined with various base DocRE models. Experimental results show that SIEF consistently improves the performance of base models in different domains, and that it improves the robustness of DocRE models.

Acknowledgments

We are grateful to the anonymous reviewers and meta reviewers for their insightful comments and suggestions. This work is funded by the National Key Research and Development Program of China (No. 2020AAA0108000). The corresponding authors are Kehai Chen and Tiejun Zhao. Lili Mou is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant No. RGPIN2020-04465, the Amii Fellow Program, the Canada CIFAR AI Chair Program, a UAHJIC project, a donation from DeepMind, and Compute Canada (www.computecanada.ca).

References

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

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 4925–4936.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 837–840.

Pankaj Gupta, Subburam Rajaram, Hinrich Schütze, and Thomas A. Runkler. 2019. Neural relation extraction within and across sentence boundaries. In *Proceedings of the Association for the Advance of Artificial Intelligence*, pages 6513–6520.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *Proceedings of the Annual Conference on Neural Information Processing Systems Deep Learning and Representation Learning Workshop*.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021. Three sentences are all you need: Local path enhanced document relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 998–1004.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 1003–1011.

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

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics*, pages 113–120.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 1171–1182.

Linfeng Song, Yue Zhang, Daniel Gildea, Mo Yu, Zhiguo Wang, and Jinsong Su. 2019. Leveraging dependency forest for neural medical relation extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789.

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

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488.

Minguang Xiao and Cong Liu. 2016. Semantic relation classification via hierarchical recurrent neural network with attention. In *Proceedings of the International Conference on Computational Linguistics: Technical Papers*, pages 1254–1263.

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

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021b. Document-level relation extraction with reconstruction. In *Proceedings of the Association for the Advance of Artificial Intelligence*, 16, pages 14167–14175.

Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. 2016. Improved relation classification by deep recurrent neural networks with data augmentation. In *Proceedings of the International Conference on Computational Linguistics: Technical Papers*, pages 1461–1470.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1785–1794.

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 Annual Meeting of the Association for Computational Linguistics*, 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 Conference on Empirical Methods in Natural Language Processing*, pages 7170–7186.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344.

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

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*, pages 3999–4006.

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 Association for the Advance of Artificial Intelligence*, 16, pages 14612–14620.

Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. 2019. Graph neural networks with generated parameters for relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1331–1339.