On the Robustness of Document-Level Relation Extraction Models to Entity Name Variations

Anonymous ACL submission

Abstract

Driven by the demand for cross-sentence and large-scale relation extraction, document-level relation extraction (DocRE) has attracted increasing research interest. Despite the continuous improvement in performance, we find that existing DocRE models which initially perform well may make more mistakes when merely changing the entity names in the document, hindering the generalization to novel entity names. To this end, we systematically investigate the robustness of DocRE models to entity name variations in this work. We first propose a principled pipeline to generate entity-renamed documents by replacing the original entity names with names from Wikidata. By applying the pipeline to DocRED and Re-DocRED datasets, we construct two novel benchmarks named Env-DocRED and Env-Re-DocRED for robustness evaluation. Experimental results show that both three representative DocRE models and two LLM-based in-context learning methods consistently lack sufficient robustness to entity name variations. Finally, we propose an entity variation robust training method which not only effectively improves the robustness of DocRE models but also enhances their understanding and reasoning capabilities.

1 Introduction

The demand for cross-sentence and large-scale relation extraction has led to a surge of research interest in document-level relation extraction (DocRE), which aims to identify the relations between each pair of entities within a document (Yao et al., 2019). While covering more realistic scenarios than its sentence-level counterpart, DocRE also brings new challenges, requiring a comprehensive modeling of interactions among different mentions of an entity, different entities and different entity pairs.

Recently, a series of DocRE studies propose various novel models and methods, continuously improving the performance on several DocRE benchmarks (Tan et al., 2022a; Zhou and Lee, 2022; Xiao et al., 2022; Sun et al., 2023). However, we observe that existing DocRE models may produce more erroneous predictions when we merely change the entity names in a test document. As illustrated in Figure 1, a well-trained DocRE model correctly extracts all four relation instances from the original document. However, after replacing the entity names in the document with a new set of names of the same entity types (e.g., change the song name `Uptown Girl` into another song name `Endless Love`), the model starts making mistakes, including both false positive and false negative predictions. This indicates that existing DocRE models may overly rely on entity information for extraction and lack robustness. Considering the vast and diverse space of entity names in real-world scenarios, which also expands constantly with numerous novel entity names, the poor robustness and gen-

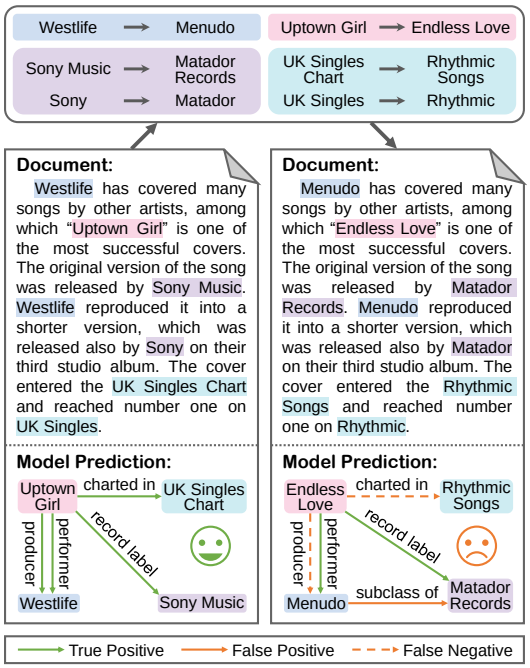

Figure 1: An illustration of how minor changes in entity names mislead the DocRE model to wrong predictions.

eralization further impedes the reliable application of DocRE models.

As a result, we systematically study the robustness of DocRE models to entity name variations in this work. To audit the robustness of existing DocRE models, we first propose a general pipeline to automatically generate perturbed test documents with changed entity names. Building such a pipeline is non-trivial for three reasons: (1) the relation types are constrained by entity types, for instance, the tail entity of relation `record label` in Figure 1 must be a record label, therefore the new entity name should not alter the original entity type, otherwise the relation labels may no longer hold; (2) for an entity mentioned multiple times under different names, each alias should be replaced with a distinct name to exclude the interference caused by different coreference structures, like `Sony Music` ⇒ `Matador Records` and `Sony` ⇒ `Matador` in Figure 1; (3) the introduced entity names should be of high quality and come from a wide range of sources. We strictly adhere to the principles above and design a four-stage pipeline based on Wikidata, which retrieves valid items from Wikidata for entity name substitution.

We further apply the proposed pipeline to DocRED (Yao et al., 2019) and Re-DocRED (Tan et al., 2022b), due to both being the largest and most widely used DocRE datasets, to create two novel benchmarks, named Env-DocRED and Env-Re-DocRED, for evaluating the robustness of DocRE models to entity name variations[1]. By conducting extensive experiments on both original and newly constructed benchmarks, we thoroughly evaluate the robustness of three representative DocRE models and two LLM-based in-context learning method. The results show that the performance of all evaluated models drops significantly on Env-DocRED and Env-Re-DocRED (e.g., the best model's F1 drops from 79.3% on Re-DocRED to 57.0% on Env-Re-DocRED), revealing the poor robustness to entity name variations. In order to gain more in-depth insights, we also conduct detailed analyses in terms of models' performance bottleneck, robustness on intra- and inter-sentence relations, and the relationship between robustness and entity count, etc.

Finally, to improve the robustness of DocRE models to entity name variations, we propose an **E**ntity **V**ariation **R**obust **T**raining method (EVRT) which is based on data augmentation and consistency regularization. For each training document, we generate a perturbed document by entity renaming. Then, in addition to the classification loss for entity pairs in the original document, our method introduces three extra objectives, which respectively penalize the classification errors for entity pairs in the perturbed document, the inconsistency between representations, and inconsistency between predictions of original and corresponding perturbed entity pairs. Experimental results demonstrate that EVRT not only improve the robustness of existing DocRE models but also enhance their understanding and reasoning capabilities. Besides, we transfer the idea of EVRT to in-context learning and propose a simple prompt optimization strategy, which effectively enhances the robustness of in-context learning of DocRE.

2 Related Work

Document-Level Relation Extraction. Driven by the demand for cross-sentence and large-scale relation extraction, research on relation extraction has expanded from sentence level to document level (Quirk and Poon, 2017; Yao et al., 2019). Recently document-level relation extraction has attracted increasing research interest, with new models emerging constantly. Based on the way of modeling relational information from the context, existing studies can be categorized into graph-based and sequence-based approaches. The former typically abstract the document by graph structures and perform inference with graph neural networks (Zeng et al., 2020; Zhang et al., 2021a; Wei and Li, 2022; Lu et al., 2023), while the latter encode the long-distance contextual dependencies with transformer-only architectures (Zhou et al., 2021; Xie et al., 2022; Zhang et al., 2022; Ma et al., 2023).

Robustness and Entity-Related Robustness in NLP. Despite achieving great progress with large pre-trained language models in various tasks, modern NLP models are still brittle to out-of-domain data (Hendrycks et al., 2020), adversarial attacks (McCoy et al., 2019) or small perturbation to the input (Ebrahimi et al., 2018). Consequently, there has been a growing effort to explore robustness issues in NLP, such as building robustness evaluation benchmarks and proposing robustness enhancement strategies (Wang et al., 2022). One branch of works focus on entity-related robustness

[1]Our proposed pipeline can also be applied or adapted to other DocRE datasets, which we discuss in detail in Section 9.

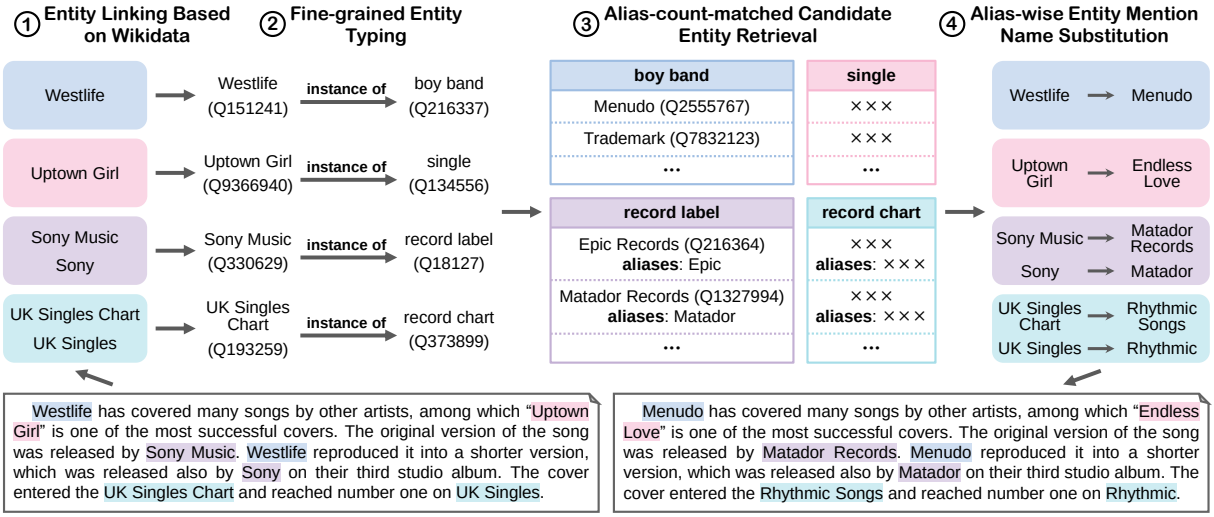

Figure 2: The proposed pipeline for generating documents with changed entity names.

of NLP models. By introducing various types of perturbations to entity (names), previous works audit or improve model robustness on different tasks like named entity recognition (Lin et al., 2021), machine reading comprehension (Yan et al., 2022) and dialogue state tracking (Cho et al., 2022).

Robustness of DocRE Models. Compared with other NLP areas, research on robustness in DocRE is relatively scarce. Xu et al. (2022) observe that DocRE models may err when non-evidence sentences of a document are removed and propose a sentence focusing loss to improve the robustness. Chen et al. (2023) reveals the poor robustness of DocRE models to word-level attacks such as synonym substitution. A few recent works also construct entity-level attacks to investigate the robustness of DocRE models (Li et al., 2023; Chen et al., 2023). However, all these attacks are not natural or adversarial, as they either disrupt entity structures (e.g., random entity mention drop) or alter entity types (e.g., random out-of-distribution entity substitution from a very limited source), rendering partial relation labels no longer valid. In contrast, we propose a principled pipeline to generate entity-renamed documents with labels preserved, and systematically evaluate and improve the robustness of DocRE models to entity name variations.

3 Problem Formulation

Given a document D which contains a set of entities $\mathcal{E} = \{e_i\}_{i=1}^{N_e}$, the task of document-level relation extraction is to predict the set of all possible relations between each entity pair $(e_h, e_t) \in \{(e_i, e_j) \mid i, j = 1, \ldots, N_e; i \neq j\}$ from a predefined relation type set $\mathcal{R}$. The subscripts of e_h and e_t refer to the head and tail entity in an entity pair. An entity e_i can occur multiple times in the document via N_{e_i} mentions $\mathcal{M}_{e_i} = \{m_j^i\}_{j=1}^{N_{e_i}}$, where the mention m_j^i refers to the token span of e_i's j-th occurrence in the document.

4 Benchmark Construction

In this section, we elaborate on the process of constructing benchmarks for evaluating the robustness of DocRE models to entity name variations. We first propose a general pipeline to generate documents with changed entity names, then apply the pipeline to DocRED and Re-DocRED to create the Env-DocRED and Env-Re-DocRED benchmarks.

4.1 Construction Pipeline

As shown in Figure 2, our proposed pipeline consists of the following four steps.

Step 1: Entity Linking Based on Wikidata. Given a document, we first link each entity in the document to an item in Wikidata. Each item in Wikidata have a label and any number of aliases, and is uniquely identified by a number starting with "Q". For example, we link the entity `Westlife` to item `Westlife(Q151241)` in Wikidata. Depending on the dataset at hand, we can perform entity linking using Wikidata Search API, off-the-shelf tools or methods specifically optimized for the datasets.

Step 2: Fine-grained Entity Typing. Next we query the value of `Instance Of` property (numbered as `P31` in Wikidata) for each linked item on Wikidata, to obtain the fine-grained

type of each entity, like boy band(Q216337) for Westlife(Q151241) in Figure 2.

Step 3: Alias-count-matched Candidate Entity Retrieval. Based on the fine-grained type of each entity, we further retrieve additional Wikidata items with the same entity type as candidates by executing a reverse query of Step 2. Note that we only retain those items whose number of aliases (plus label) are greater than or equal to the number of aliases of the original entity in the document. For example, since the entity Sony Music is mentioned under two different names in the document, we only take the retrieved items of record label with at least one Wikidata alias.

Step 4: Alias-wise Entity Mention Name Substitution. Finally, for each entity in the document, we randomly select an item from its candidate set and use this item to perform alias-wise entity mention name substitution, i.e., substitute a distinct name of the item for each alias of the original entity, like Sony Music ⇒ Matador Records and Sony ⇒ Matador in Figure 2.

4.2 Env-DocRED and Env-Re-DocRED Benchmarks

With the proposed pipeline, we further construct the robustness evaluation benchmarks based on existing datasets, which we choose DocRED (Yao et al., 2019) and Re-DocRED (Tan et al., 2022b) in this work. DocRED is one of the largest and most popular public datasets for DocRE, which is collected from English Wikipedia documents. DocRED has 96 pre-defined relation types, along with 3053/1000/1000 documents for training/development/test. Each document in DocRED has 19.5 entities and 12.5 relation triples on average. Re-DocRED is a revised version of DocRED, resolving the missing relation issue in DocRED. The 3053 revised training documents contain 28.1 triples on average and 1000 revised development documents (split into 500/500 development/test documents) have 34.7 triples on average.

We iterate over the development and test set of DocRED and Re-DocRED and apply the pipeline five times on each document with different random seeds. We name the two newly constructed benchmarks Env-DocRED and Env-Re-DocRED, with the former having 3053/5000/5000 and the latter having 3053/2500/2500 documents for training/development/test. We adopt the entity linking method and results of Genest et al. (2023) in Step 1, which has a high quality benefited from its specific design for DocRED. Besides, since all entities of NUM ant TIME type in (Re-)DocRED can not be linked to Wikidata, we take a rule-based substitution method to produce novel names for number and time. Although a small portion of entities remain unlinked, statistics show that we have altered the names of over 92% entities in original datasets.

5 Robustness Evaluation and Analysis

In this section, we conduct a comprehensive robustness evaluation and analysis on three representative DocRE models: DocuNet (Zhang et al., 2021b), KDDocRE (Tan et al., 2022a) and NCRL (Zhou and Lee, 2022) (refer to Appendix A for more details on models and implementations).

5.1 Main Evaluation Results

We present the evaluation results on the test sets of four benchmarks in Table 3. We can observe that all DocRE models have a significant performance fluctuations on Env-DocRED and Env-Re-DocRED, with the relative F1 drop ranging from 21% ~31%, revealing the insufficient robustness to entity name variations. Model-wise, we find that the three selected DocRE models show similar relative decline in performance, with none being significantly more robust than others. Encoder-wise, we find that RoBERTa$_{\text{large}}$ with higher performance also leads to better robustness than BERT$_{\text{base}}$. Dataset-wise, somewhat surprisingly, the relative decrease in F1 is even larger on Env-Re-DocRED than Env-DocRED. This suggests that despite Re-DocRED providing more complete relation labels, DocRE models still fail to gain benefits in robustness.

5.2 Further Analysis

To gain deeper insights, we conduct further analysis by answering the following questions.

Q1: What is the performance bottleneck of DocRE models under entity name variations?

Given that the entity name variations lead to a drop in performance, a natural question is whether the model generate more false positive or false negative predictions. To better understand the performance bottleneck of DocRE models, we compare the changes in precision and recall of three models with BERT$_{\text{base}}$ encoder. As shown in Table 1, the recall across models decreases significantly, while the precision changes little and even increases on

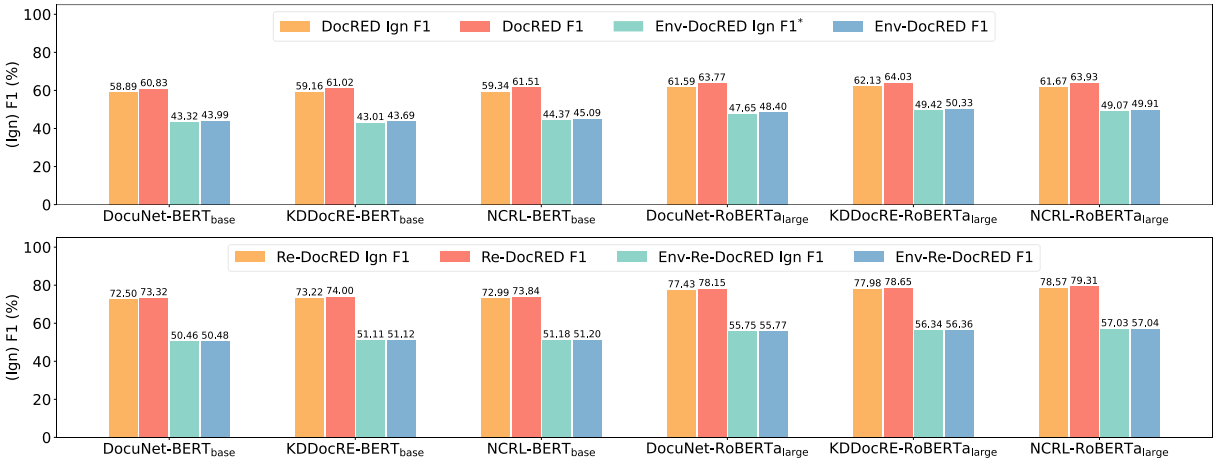

Figure 3: Evaluation results on the test sets of four benchmarks. Since the test set of DocRED is unpublished, the Ign F1 results on Env-DocRED are not accurate and marked with "*", same applies to Table 6.

Model	DocRED		Env-DocRED		Re-DocRED		Env-Re-DocRED	
	P	R	P	R	P	R	P	R
DocuNet	62.88	58.67	64.56	33.23	84.21	64.93	82.05	36.45
KDDocRE	63.95	58.76	64.27	33.61	85.04	65.51	81.50	37.24
NCRL	63.62	59.08	65.69	34.50	84.64	65.50	81.53	37.32

Table 1: Precision and recall results on the development sets of (Env-)DocRED and test sets of (Env-)Re-DocRED, same choices apply to Table 2, Figure 4 and Table 7.

Model	DocRED		Env-DocRED		Re-DocRED		Env-Re-DocRED	
	Intra	Inter	Intra	Inter	Intra	Inter	Intra	Inter
DocuNet	66.99	53.11	52.76	31.34	76.05	70.92	58.75	42.27
KDDocRE	67.33	54.03	53.12	31.64	76.89	71.40	59.48	42.81
NCRL	67.47	53.84	54.20	32.58	76.44	71.57	59.86	42.51

Table 2: Intra and Inter F1 results on four benchmarks.

Env-DocRED. This indicates that false negative predictions dominates the poorer robustness to entity name variations.

Q2: Do models show poorer robustness when predicting inter-sentence relations?

Since a major feature of DocRE is to extract the complex cross-sentence relations, we further analyse models' robustness in predicting intra-sentence and inter-sentence relations. We report the Intra F1 and Inter F1 of three $\text{BERT}_{\text{base}}$ encoded DocRE models in Table 2, which respectively evaluate on the entity pairs with and without mentions in same sentence. We can observe that on both Env-DocRED and Env-Re-DocRED, the relative F1 drop for inter-sentence relations is approximately twice that of intra-sentence relations, which indicates that existing DocRE models show poorer robustness to entity name variations when predicting inter-sentence relations.

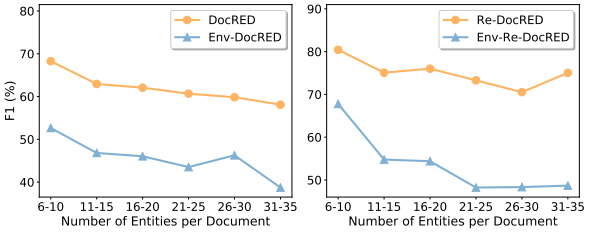

Figure 4: F1 score of $\text{NCRL-BERT}_{\text{base}}$ on documents with different number of entities.

Q3: How does the model robustness vary with the number of entities in the document?

We also investigate the robustness of DocRE models on documents with varying number of entities. This aids in better extrapolating our findings to longer documents, which often contain more entities. We divide the documents into different groups by the number of entities and evaluate the performance on each group. We showcase the results of $\text{NCRL-BERT}_{\text{base}}$ in Figure 4. As the number of entities increases, the absolute performance drop under entity name variations gets larger, especially on Env-Re-DocRED. The slopes of the linear fits on DocRED, Env-DocRED, Re-DocRED, Env-Re-DocRED are -0.35, -0.42, -0.24 and -0.69 respectively. Note that the performance itself also shows a decreasing trend when encountering more entities, thus the relative performance drop should be more significant. This suggests that the model may be more brittle as the number of entities increases.

Q4: How can we disentangle the reasons for the performance drop?

Yan et al. (2022) pointed out that the information associated with the entity name that can be lever-

Type	DocRED			Env-DocRED			Re-DocRED			Env-Re-DocRED		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
PER	32	68	161	7	11	19	33	65	155	8	12	19
ORG	27	104	587	7	12	27	34	125	685	7	13	28
LOC	27	128	1240	8	20	90	29	148	1704	8	21	108
MISC	17	37	141	7	12	23	18	42	171	7	12	23
Total	25	73	309	7	13	29	27	81	393	7	13	32

Table 3: The upper quartile (Q3), median (Q2) and lower quartile (Q1) of entity popularities of four benchmarks' test sets (only calculating entities with name changed, same applies to Table 4).

Type	DocRED	Env-DocRED	Re-DocRED	Env-Re-DocRED
PER	12.33%	1.90%	12.72%	2.23%
ORG	25.35%	3.47%	28.21%	3.14%
LOC	32.85%	8.25%	37.69%	10.77%
TIME	34.02%	16.62%	41.62%	20.82%
NUM	34.74%	12.01%	41.78%	16.86%
MISC	18.11%	3.04%	19.71%	3.11%
Total	23.47%	5.28%	27.34%	6.89%

Table 4: The proportion of entity mentions that appear in training sets of four benchmarks' test sets.

aged by the model includes both entity knowledge and name clues. The former refers to the world knowledge associated with the entity like "Westlife is a famous boy band", which mainly comes from pre-training. The latter refer to the statistical clues associated with the name's surface form like "Westlife always appears with the performer relation in training set", which mainly comes from fine-tuning. The perturbations to entity names may break these two types of information.

We adopt two measurements to better understand the information loss. We calculate the popularity of entities (Huang et al., 2022), i.e., how many times the linked item of the entity appears in a relation instance in Wikidata, in each benchmark's test set to roughly quantify the entity knowledge. As shown in Table 3, the popularity of entities in two new benchmarks drops significantly. For name clues, we calculate the percentage of entity mentions that appear in training sets for each benchmark's test set. As shown in Table 4, the proportion also have a noticeable drop in two novel benchmarks.

Q5: How robust is the in-context learning of LLMs under entity name variations?

Recently large language models (LLM) (Brown et al., 2020) have achieved promising few-shot results on many tasks through in-context learning (ICL) (Dong et al., 2023). Therefore, we also conduct an experiment to explore how robust of ICL for DocRE under entity name variations. We use

Model	Re-DocRED		Env-Re-DocRED	
	1-Shot	3-Shot	1-Shot	3-Shot
GPT-3.5 Turbo	13.66	16.00	10.81	12.98
GPT-4 Turbo	28.35	32.41	21.59	23.08

Table 5: F1 score of LLM-based ICL DocRE methods on the test sets of Re-DocRED and Env-Re-DocRED.

gpt-3.5-turbo-0125[2] and gpt-4-0125-preview[3] due to them being the most capable LLMs currently. We experiment on both 1-Shot and 3-Shot settings, which represent providing 1 and 3 example document(s) and gold relation instances as demonstrations. We randomly select demonstration document in the training set for each test document and set the temperature parameter to 0 for least randomness. The experimental results on test sets of Re-DocRED and Env-Re-DocRED are shown in Table 5. We find that on both settings, the two LLM-based ICL approaches have a performance drop on Env-Re-DocRED, suggesting that the robustness issue exists not only in specialized models but also in large models.

6 Entity Variation Robust Training

Due to the unsatisfactory robustness of existing DocRE models to entity name variations, we further explore the method for enhanced robustness. Intuitively, we can adopt a similar approach as the proposed pipeline to perturb each training document with a group of new entity names. The derived document naturally shares the same relation labels with the original one. Also, a robust DocRE models should generate consistent representations and predictions for each corresponding entity pair in the original and perturbed documents. Based on such motivation, we propose an entity variation robust training method (EVRT) that is enhanced by data augmentation and consistency regularization.

Specifically, given a labeled entity pair (e_h, e_t) in a document, vanilla approaches typically train the DocRE model with a classification objective $\mathcal{L}_{clo} = \ell_{task}(e_h, e_t)$, where ℓ_{task} denotes the loss function depending on the specific model.

Denoting the corresponding entity pair of (e_h, e_t) in the perturbed document as $(e_{\hat{h}}, e_{\hat{t}})$, our proposed method first incorporate the classification loss $\mathcal{L}_{clp} = \ell_{task}(e_{\hat{h}}, e_{\hat{t}})$ for $(e_{\hat{h}}, e_{\hat{t}})$ to penalize the

[2]https://platform.openai.com/docs/models/gpt-3-5-turbo

[3]https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo (Due to limited budget, the experiments with gpt-4-0125-preview only use 1/5 documents.)

classification errors for entity pairs in the perturbed document. Then we introduce representation consistency regularization and prediction consistency regularization to encourage the model to produce consistent representations and predicted probability distributions between (e_h, e_t) and $(e_{\hat{h}}, e_{\hat{t}})$. Formally, we define the representation consistency regularization loss as:

$$\mathcal{L}_{rcr} = \|\boldsymbol{z}^{(h,t)} - \boldsymbol{z}^{(\hat{h},\hat{t})}\|_2^2, \quad (1)$$

where $\boldsymbol{z}^{(h,t)}$ is the pair representation of (e_h, e_t). And we define the prediction consistency regularization loss as:

$$\mathcal{L}_{pcr} = \sum_{r \in \mathcal{R}} \mathcal{D}_{SKL}(\boldsymbol{p}_r^{(h,t)}, \boldsymbol{p}_r^{(\hat{h},\hat{t})}), \quad (2)$$

where $\boldsymbol{p}_r^{(h,t)} = [P_r^{(h,t)}, 1 - P_r^{(h,t)}]$, $P_r^{(h,t)}$ is the predicted probability of relation r for (e_h, e_t), $\mathcal{D}_{SKL}$ is the symmetric KL divergence:

$$\mathcal{D}_{SKL}(\boldsymbol{p}, \boldsymbol{q}) = \mathcal{D}_{KL}(\boldsymbol{p}\|\boldsymbol{q}) + \mathcal{D}_{KL}(\boldsymbol{q}\|\boldsymbol{p}), \quad (3)$$

where $\mathcal{D}_{KL}$ is the vanilla KL divergence. The overall objective is defined as:

$$\mathcal{L} = \mathcal{L}_{clo} + \mathcal{L}_{clp} + \alpha\mathcal{L}_{rcr} + \beta\mathcal{L}_{pcr}, \quad (4)$$

where α and β are two hyperparameters. Note that to prevent the incorporated novel entity names for training document perturbation have overlap with those entity names for substitution when constructing the benchmarks, resulting in potential shortcuts, we isolate the new entity names introduced during benchmark construction when replacing the entities in training documents.

7 Experiments

7.1 Main Results

The main results on the test sets of four benchmarks are shown in Table 6. It is shown that when equipped with the proposed EVRT method, all DocRE models achieve a significant performance gain on Env-DocRED (a maximum more then 9% absolute increase in F1) and Env-Re-DocRED (a maximum more than 12% absolute increase in F1). Meanwhile, the performance on DocRED and Re-DocRED only shows a slight drop. All these results indicate that EVRT can effectively improve the robustness of existing DocRE models to entity name variations.

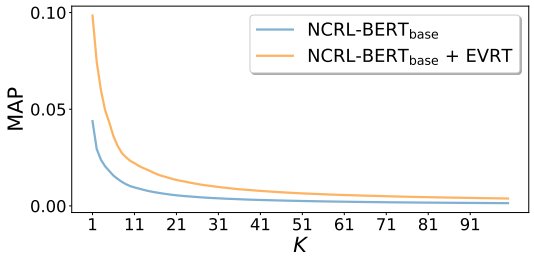

Figure 5: MAP curves of NCRL-BERT$_{\text{base}}$ and NCRL-BERT$_{\text{base}}$ + EVRT.

7.2 Ablation Study

We further conduct an ablation study on Env-DocRED and Env-Re-DocRED to investigate the influence of three newly added training objective. As shown in Table 7, only introducing one of $\mathcal{L}_{clp}$, $\mathcal{L}_{rcr}$ and $\mathcal{L}_{pcr}$ has lead to a significant performance improvement, which indicates the effectiveness of each objective. When combining these losses pairwise, the performance is further enhanced. And the best performance is achieved when simultaneously using three objectives together. We also observe that compare to $\mathcal{L}_{rcr}$, $\mathcal{L}_{clp}$ and $\mathcal{L}_{pcr}$ may play a more important role for the improvement.

7.3 Understanding and Reasoning Capability Evaluation

We also take the MAP evaluation metric proposed in Chen et al. (2023) to evaluate the understanding and reasoning capabilities of the DocRE models trained with and without our EVRT method. Given top K words with the highest attribution values, the formula of MAP over T relational facts is:

$$\text{MAP}(K) = \frac{1}{T}\sum_{t=1}^{T}\text{AP}_t(K) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{K}\sum_{i=1}^{K}P_t(i)\cdot\mathbf{1}_t(i), \quad (5)$$

where $\mathbf{1}_t(i)$ is the indicator function of the i-th important word for predicting the t-th relational fact. We select all possible values of K an report the MAP curve of NCRL-BERT$_{\text{base}}$ and NCRL-BERT$_{\text{base}}$ + EVRT models in Figure 5. It is observed that the MAP values of NCRL-BERT$_{\text{base}}$ + EVRT are consistencly higher than NCRL-BERT$_{\text{base}}$, suggesting that the proposed EVRT method not only improves the robustness of DocRE models but also enhances their understanding and reasoning capabilities.

Model	DocRED		Env-DocRED		Re-DocRED		Env-Re-DocRED	
	Ign F1	F1	Ign F1*	F1	Ign F1	F1	Ign F1	F1
DocuNet-BERT$_{\text{base}}$	58.89	60.83	43.32	43.99	72.50$_{\pm 0.17}$	73.32$_{\pm 0.20}$	50.46$_{\pm 0.44}$	50.48$_{\pm 0.44}$
+ **EVRT**	58.17 (↓ 0.72)	59.71 (↓ 1.12)	51.63 (↑ 8.31)	52.78 (↑ 8.79)	71.64$_{\pm 0.12}$ (↓ 0.86)	72.44$_{\pm 0.19}$ (↓ 0.88)	62.32$_{\pm 0.46}$ (↑ 11.86)	62.33$_{\pm 0.46}$ (↑ 11.85)
KDDocRE-BERT$_{\text{base}}$	59.16	61.02	43.01	43.69	73.22$_{\pm 0.27}$	74.00$_{\pm 0.30}$	51.11$_{\pm 0.58}$	51.12$_{\pm 0.58}$
+ **EVRT**	58.69 (↓ 0.47)	60.21 (↓ 0.81)	51.64 (↑ 8.63)	52.94 (↑ 9.25)	72.41$_{\pm 0.18}$ (↓ 0.81)	73.25$_{\pm 0.15}$ (↓ 0.75)	62.53$_{\pm 0.19}$ (↑ 11.42)	62.55$_{\pm 0.19}$ (↑ 11.43)
NCRL-BERT$_{\text{base}}$	59.34	61.51	44.37	45.09	72.99$_{\pm 0.28}$	73.84$_{\pm 0.32}$	51.18$_{\pm 0.62}$	51.20$_{\pm 0.62}$
+ **EVRT**	58.84 (↓ 0.50)	60.51 (↓ 1.00)	52.97 (↑ 8.60)	54.25 (↑ 9.16)	72.00$_{\pm 0.36}$ (↓ 0.99)	72.78$_{\pm 0.42}$ (↓ 1.06)	62.83$_{\pm 0.25}$ (↑ 11.65)	62.84$_{\pm 0.25}$ (↑ 11.64)
DocuNet-RoBERTa$_{\text{large}}$	61.59	63.77	47.65	48.40	77.43$_{\pm 0.26}$	78.15$_{\pm 0.25}$	55.75$_{\pm 0.70}$	55.77$_{\pm 0.70}$
+ **EVRT**	60.48 (↓ 1.11)	62.46 (↓ 1.31)	54.32 (↑ 6.67)	55.93 (↑ 7.53)	76.07$_{\pm 0.14}$ (↓ 1.36)	76.68$_{\pm 0.18}$ (↓ 1.47)	67.37$_{\pm 0.27}$ (↑ 11.62)	67.38$_{\pm 0.27}$ (↑ 11.61)
KDDocRE-RoBERTa$_{\text{large}}$	62.13	64.03	49.42	50.33	77.98$_{\pm 0.22}$	78.65$_{\pm 0.23}$	56.34$_{\pm 0.61}$	56.36$_{\pm 0.61}$
+ **EVRT**	60.49 (↓ 1.64)	62.20 (↓ 1.83)	56.50 (↑ 7.08)	57.83 (↑ 7.50)	76.20$_{\pm 0.41}$ (↓ 1.78)	76.82$_{\pm 0.43}$ (↓ 1.83)	68.60$_{\pm 0.25}$ (↑ 12.26)	68.62$_{\pm 0.25}$ (↑ 12.26)
NCRL-RoBERTa$_{\text{large}}$	61.67	63.93	49.07	49.91	78.57$_{\pm 0.22}$	79.31$_{\pm 0.26}$	57.03$_{\pm 0.94}$	57.04$_{\pm 0.94}$
+ **EVRT**	60.28 (↓ 1.39)	62.21 (↓ 1.72)	56.29 (↑ 7.22)	57.81 (↑ 7.90)	76.78$_{\pm 0.19}$ (↓ 1.79)	77.48$_{\pm 0.21}$ (↓ 1.83)	68.87$_{\pm 0.19}$ (↑ 11.84)	68.89$_{\pm 0.19}$ (↑ 11.85)

Table 6: Main results on the test sets of four benchmarks.

$\mathcal{L}_{clp}$	$\mathcal{L}_{rcr}$	$\mathcal{L}_{pcr}$	Env-DocRED		Env-Re-DocRED	
			Ign F1	F1	Ign F1	F1
–	–	–	45.21	45.23	51.18	51.20
✓	–	–	52.89	52.91	62.05	62.06
–	✓	–	52.13	52.14	61.08	61.10
–	–	✓	53.36	53.38	61.83	61.84
✓	✓	–	52.75	52.77	62.21	62.22
✓	–	✓	53.79	53.80	62.41	62.42
–	✓	✓	53.50	53.52	62.09	62.11
✓	✓	✓	**54.15**	**54.17**	**62.83**	**62.84**

Table 7: Ablation study results.

Model	Re-DocRED		Env-Re-DocRED	
	1-Shot	3-Shot	1-Shot	3-Shot
GPT-3.5 Turbo	13.66	16.00	10.81	12.98
+ DA	14.67	16.47	11.59	13.86
+ DA + CG	**15.14**	**17.22**	**12.44**	**14.37**
GPT-4 Turbo	28.35	32.41	21.59	23.08
+ DA	28.20	33.52	22.85	24.41
+ DA + CG	**28.99**	**34.32**	**23.74**	**25.11**

Table 8: F1 score of entity variation robust in-context learning method for DocRED.

7.4 Entity Variation Robust In-Context Learning

The results in Section 5.2 indicates that utilize in-context learning of LLMS for DocRE also shows insufficient robustness to entity name variations. A natural question is can we transfer the basic idea of EVRT to improve the robustness of in-context learning. We conduct a preliminary attempt by designing a simple entity variation robust in-context learning method, which optimize the prompt with demonstration augmentation (DA) and consistency guidance (CG). Based on the vanilla prompts, demonstration augmentation add an entity-renamed document for each original demonstration document. And consistency guidance further expand the prompt by explicitly explains that "each pair of original and augmented demonstration documents only differs in entity names and thus have consistent relation labels" and "please take the consistency into consideration for better predictions". As shown in Table 8, this simple strategy also effectively enhances the robustness of LLM-based in-context learning methods.

8 Conclusion

Our main contributions in this work are three-fold: (1) Resource-wise, we propose a general pipeline to reasonably generate entity-renamed documents and construct two novel benchmarks, Env-DocRED and Env-Re-DocRED, for robustness evaluation. (2) Experiment-wise, we conduct comprehensive experiments on multiple DocRE models to evaluate their robustness and provide further analyses from multiple perspectives. (3) Methodology-wise, we propose entity variance robust training and in-context learning methods, effectively improving the robustness of DocRE models.

9 Limitations and Future Directions

In this section, we analyse the limitations of our work from three perspectives and hope to provide inspiration for future works.

Task Setting. Our study is grounded upon a classic setting of DocRE where the entity information including entity mention positions and coreference clusters of mentions are given beforehand. Some recent works explore the end-to-end setting of DocRE, which requires the model to jointly perform mention detection (and optionally classification), coreference resolution and relation extraction, aligning better with real-world application scenarios (Eberts and Ulges, 2021; Xu and Choi, 2022; Zhang et al., 2023). Investigating the robustness of end-to-end DocRE approaches to entity name variations is a promising direction for future works. More importantly, since the proposed pipeline for entity name substitution does not alter entity types and coreference labels, our constructed benchmarks can be directly utilized for the study of end-to-end DocRE model robustness, rendering the two benchmarks more valuable.

Dataset Domain and Language. Given that we construct the robustness evaluation benchmarks based on DocRED and Re-DocRED, which originate from English Wikipedia documents, our findings may be somewhat limited to English, generic-domain scenarios. Leveraging other well-established DocRE datasets, future works are encouraged to extend the study on entity name variation robustness of DocRE models to more domains such as news (Zaporojets et al., 2021), biomedicine (Li et al., 2016) and scientific publications (Luan et al., 2018), and more languages such as Chinese (Cheng et al., 2021) and Korean (Yang et al., 2023). As Wikidata covers a wide range of domains and languages, the proposed benchmark construction pipeline can also be applied to other datasets. For datasets that are hard to be linked to Wikidata, one may explore the possibility of adapting the pipeline with an appropriate knowledge base.

Methodology. Since the proposed entity variance robust training and in-context learning frameworks generate a perturbed document with entity names changed for each training document, fine-tuning pre-trained models incurs larger memory overhead, and utilizing large language models for in-context learning entails higher time and cost expenses. Additionally, although the proposed methods significantly improve the performance of multiple models on Env-DocRED and Env-Re-DocRED, there is still a certain gap compared to DocRED and Re-DocRED. An intriguing avenue for future research is to explore more efficient and effective techniques to improve the robustness of DocRE models to entity name variations.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901.

Haotian Chen, Bingsheng Chen, and Xiangdong Zhou. 2023. Did the models understand documents? benchmarking models for language understanding in document-level relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6418–6435.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2819–2831.

Hyundong Cho, Chinnadhurai Sankar, Christopher Lin, Kaushik Sadagopan, Shahin Shayandeh, Asli Celikyilmaz, Jonathan May, and Ahmad Beirami. 2022. Know thy strengths: Comprehensive dialogue state tracking diagnostics. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5345–5359.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

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: Main Volume*, pages 3650–3660.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

Pierre-Yves Genest, Pierre-Edouard Portier, Elöd Egyed-Zsigmond, and Martino Lovisetto. 2023. Linked-docred - enhancing docred with entity-linking to evaluate end-to-end document-level information extraction pipelines. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 3064–3074.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

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 (Volume 1: Long Papers)*, pages 6241–6252.

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*, page baw068.

Jing Li, Yequan Wang, Shuai Zhang, and Min Zhang. 2023. Rethinking document-level relation extraction: A reality check. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5715–5730.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. RockNER: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang, and Yongyi Mao. 2023. Anaphor assisted document-level relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15453–15464.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

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*, pages 1971–1983.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

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: Volume 1, Long Papers*, pages 1171–1182.

Qi Sun, Kun Huang, Xiaocui Yang, Pengfei Hong, Kun Zhang, and Soujanya Poria. 2023. Uncertainty guided label denoising for document-level distant relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15960–15973.

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 the Association for Computational Linguistics: ACL 2022*, pages 1672–1681.

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*, pages 8472–8487.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586.

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*, page 2000–2008.

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*, pages 2395–2409.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268.

Liyan Xu and Jinho Choi. 2022. Modeling task interactions in document-level joint entity and relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5409–5416.

Wang Xu, Kehai Chen, Lili Mou, and Tiejun Zhao. 2022. Document-level relation extraction with sentences importance estimation and focusing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2920–2929.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. On the robustness of reading comprehension models to entity renaming. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–520.

Soyoung Yang, Minseok Choi, Youngwoo Cho, and Jaegul Choo. 2023. HistRED: A historical document-level relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3207–3224.

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*, pages 764–777.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. Dwie: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, page 102563.

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, 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.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021a. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021b. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006.

Ruoyu Zhang, Yanzeng Li, and Lei Zou. 2023. A novel table-to-graph generation approach for document-level joint entity and relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10853–10865.

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*, pages 14612–14620.

Yang Zhou and Wee Sun Lee. 2022. None class ranking loss for document-level relation extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4538–4544.

A Selected Models and Evaluation Metrics

We choose three public-available DocRE models which are representative for their strong performance and high popularity. **DocuNet** (Zhang et al., 2021b) formulates DocRE as a semantic segmentation task and captures both local context information and global interdependency among triples for extraction. **KDDocRE** (Tan et al., 2022a) uses an axial attention module for two-hop relations reasoning and an adaptive focal loss to address the class imbalance problem. **NCRL** (Zhou and Lee, 2022) shares same model with a strong DocRE baseline ATLOP (Zhou et al., 2021) but improves upon the learning of none class. We use Ign F1 and F1 scores as the evaluation metrics, where Ign F1 measures the F1 excluding those relational facts shared by the training and development/test sets. For each model, we all experiment with $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) and $\text{RoBERTa}_{\text{large}}$ (Liu et al., 2019) encoder, leading to six submodels. We reimplement all models with their official codes and

report the the mean and standard deviation results by five trials with different random seeds. Since the test set of DocRED is released by Codalab, we report the official test score of the best checkpoint on development set.