Rethinking Document-Level Relation Extraction: A Reality Check

Jing Li[1], **Yequan Wang**[2], **Shuai Zhang**[3] **and Min Zhang**[1*]
[1]Harbin Institute of Technology, Shenzhen, China
[2]Beijing Academy of Artificial Intelligence, Beijing, China
[3]ETH Zurich, Switzerland
{li.jing, zhangmin2021}@hit.edu.cn,
tshwangyequan@gmail.com, cheungdaven@gmail.com

Abstract

Recently, numerous efforts have continued to push up performance boundaries of document-level relation extraction (DocRE) and have claimed significant progress in DocRE. In this paper, we do not aim at proposing a novel model for DocRE. Instead, we take a closer look at the field to see if these performance gains are actually true. By taking a comprehensive literature review and a thorough examination of popular DocRE datasets, we find that these performance gains are achieved upon a strong or even untenable assumption in common: all named entities are perfectly localized, normalized, and typed in advance. Next, we construct four types of entity mention attacks to examine the robustness of typical DocRE models by behavioral probing. We also have a close check on model usability in a more realistic setting. Our findings reveal that most of current DocRE models are vulnerable to entity mention attacks and difficult to be deployed in real-world end-user NLP applications. Our study calls more attentions for future research to stop simplifying problem setups, and to model DocRE in the wild rather than in an unrealistic Utopian world.

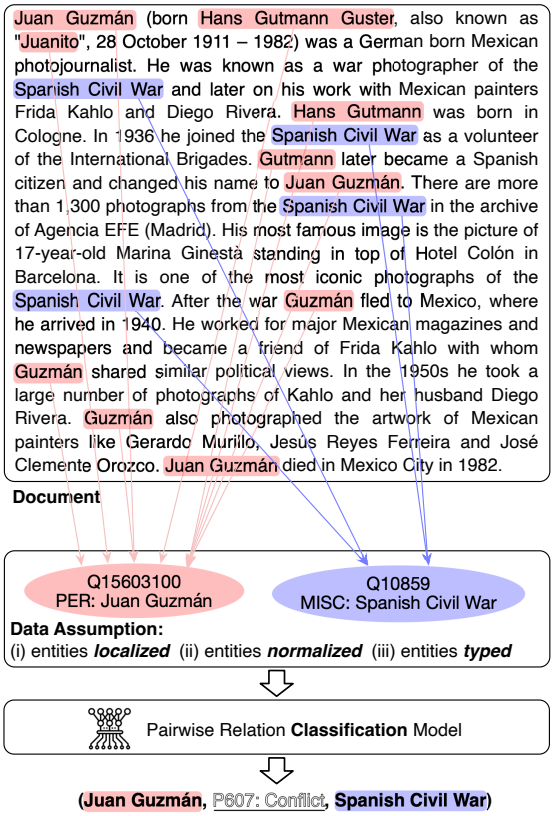


Figure 1: Data assumption in most of DocRE models.

1 Introduction

Document-level relation extraction (DocRE), aiming at identifying semantic relations between a head entity and a tail entity in a document (Yao et al., 2019), plays an essential role in a variety of downstream applications, such as question answering (Xu et al., 2016) and knowledge base construction (Trisedya et al., 2019).

Recently, there are two flourishing branches for DocRE. First, graph-based approaches consider entities (Velickovic et al., 2018; Nan et al., 2020), mentions (Christopoulou et al., 2019; Li et al., 2020) and sentences (Xu et al., 2021c) as nodes to construct a document-level graph and perform reasoning through some advanced neural graph techniques. Second, sequence-based approaches leverage BiLSTM (Huang et al., 2021; Li et al., 2021b) or Transformers (Tan et al., 2022; Zhong and Chen, 2021; Zhou and Chen, 2022) as encoders to learn document-level representations. However, all these models have one thing in common that they are based on **a strong or even untenable assumption** as shown in Figure 1: all entity mentions are (i) correctly localized; (ii) perfectly normalized; (iii) correctly typed.[1] Then, the task of modeling DocRE is usually simplified as a pairwise **classification** problem.

* Corresponding author.

[1]More illustrating examples can be found in Appx. §A

Findings of the Association for Computational Linguistics: ACL 2023, pages 5715–5730
July 9-14, 2023

Although these pairwise classification approaches have claimed significant progress in DocRE performance, we are still interested in taking a closer look at the field to see if this is actually true. In particular, many research papers have reported very decent leaderboard scores for the DocRE task. Does this mean the task of DocRE has been almost completely solved? Can the current approaches be widely used in real-world DocRE scenarios?

To answer these questions, we first take a closer look at data annotations in commonly-used DocRE datasets to check the strong data assumption (§3). We focus specifically on the annotations of named entity recognition (NER) and normalization (*i.e.,* entity linking) in detecting relations. By answering three research questions (**RQ1-3**), we find that current problem setups for DocRE are greatly simplified and unrealistic.

If the data assumption is too strict, it is not clear whether current DocRE models are robust in a variety of loose assumptions. Therefore, we construct four types of attacks regarding entity mention annotations to investigate the model robustness (§4, **RQ4**) using behavioral probing (Lasri et al., 2022; Chen et al., 2022).

To further have a look at the limitations of data assumptions, it is important to investigate the usability of existing DocRE models in real-world scenarios. Hence, we examine the capability of widely-used NER systems and entity linking systems on preparing model input formats from raw text for DocRE model deployment (§5, **RQ5**). Finally, we discuss our empirical findings and call special attentions for future research in developing DocRE models (§6).

In short, our contributions and findings are:

- We present a comprehensive literature review on recent advances for DocRE and identify a strong or even untenable assumption in modeling DocRE.
- We take a thorough examination of data annotation on three popular DocRE datasets. Detecting relations in text commonly involves multiple mentions and aliases of paired entities (*i.e.,* head and tail entities) which are currently assumed to be perfectly typed, localized and normalized before modeling DocRE.
- We construct four types of entity mention attacks to check the robustness for typical DocRE models. Most of current DocRE models are vulnerable to mention attacks (F1 drops from 7.93% to 85.51%).
- We have a close check on the usability of typical DocRE models. Under the identified data assumption, current DocRE models are very difficult to be deployed in real-world end-user NLP applications because of the need of input preparation for each pipeline module (*i.e.,* the reproduction rate of input format is only from 34.3% to 58.1%).
- We discuss our findings, and call attentions for future research to stop simplifying problem setups, and to model DocRE in the wild rather than in an unrealistic Utopian world.

2 A Quick Literature Review

In this section, we have a quick literature review of DocRE models to shed light on a global review for recent evolutions. Table 1 summarizes recent studies in anti-chronological order.

Graph-based Approaches. Graph-based approaches first construct a document-level homogeneous graph where words (Zhang et al., 2020), mentions (Christopoulou et al., 2019), entities (Zhou et al., 2020), sentences (Li et al., 2020; Xu et al., 2021a) or meta dependency paths (Nan et al., 2020) are considered as nodes and some semantic dependencies (*e.g.,* mention-mention (Christopoulou et al., 2019), mention-entity (Zeng et al., 2020), mention-sentence (Wang et al., 2020), entity-sentence (Li et al., 2020), sentence-sentence (Wang et al., 2020; Xu et al., 2021b), sentence-document (Zeng et al., 2021)) as edges. One key advantage of these approaches is that some advanced graph techniques can be used to model inter- and intra-entity interactions and perform multi-hop reasoning.

Sequence-based Approaches. Instead of introducing complex graph structures, some approaches typically model a document as a sequence of tokens and leverage BiLSTM (Huang et al., 2021; Li et al., 2021b) or Transformers (Tan et al., 2022) as encoder to capture the contextual semantics. In particular, some studies have already contributed effort to integrating entity structures (Xu et al., 2021c), concept view (Li et al., 2021a), deep probabilistic logic (Zhang et al., 2021b), U-shaped Network (Zhang et al., 2021a), relation-specific attentions (Yu et al., 2022), logic rules (Ru et al., 2021), augmenting intermediate steps (Xiao et al., 2022), sentences importance estimation (Xu et al., 2022), evidence extraction (Xie et al., 2022) and knowledge dis-

References	Venue	Claim	Performed	Annotation Assumption			Aggregation
				Localization	Linking	Typing	
(Zhang et al., 2022)	EMNLP22	Extraction	Classification	✓	✓	✗	LogSumExp
(Xie et al., 2022)	ACL22	Extraction	Classification	✓	✓	✗	LogSumExp
(Tan et al., 2022)	ACL22	Extraction	Classification	✓	✓	✗	LogSumExp
(Xiao et al., 2022)	NAACL22	Extraction	Classification	✓	✓	✓	LogSumExp
(Xu et al., 2022)	NAACL22	Extraction	Classification	✓	✓	✓	Average
(Yu et al., 2022)	NAACL22	Extraction	Classification	✓	✓	✗	Average
(Zeng et al., 2021)	ACL21	Extraction	Classification	✓	✓	✓	Average
(Li et al., 2021b)	ACL21	Extraction	Classification	✓	✓	✓	Max-pooling
(Xu et al., 2021b)	ACL21	Extraction	Classification	✓	✓	✓	Average
(Huang et al., 2021)	ACL21	Extraction	Classification	✓	✓	✗	Average
(Makino et al., 2021)	ACL21	Extraction	Classification	✓	✓	✓	Max-pooling
(Ru et al., 2021)	EMNLP21	Extraction	Classification	✓	✓	✓	Average
(Zhang et al., 2021b)	EMNLP21	Extraction	Classification	✓	✓	✓	[CLS]
(Zhang et al., 2021a)	IJCAI21	Extraction	Classification	✓	✓	✗	LogSumExp
(Xu et al., 2021c)	AAAI21	Extraction	Classification	✓	✓	✗	Average
(Xu et al., 2021a)	AAAI21	Extraction	Classification	✓	✓	✓	Average
(Li et al., 2021a)	AAAI21	Extraction	Classification	✓	✓	✓	Average
(Zhou et al., 2021)	AAAI21	Extraction	Classification	✓	✓	✗	Average
(Nan et al., 2020)	ACL20	Extraction	Classification	✓	✓	✗	Average
(Zeng et al., 2020)	EMNLP20	Extraction	Classification	✓	✓	✓	Average
(Wang et al., 2020)	EMNLP20	Extraction	Classification	✓	✓	✓	Average
(Tran et al., 2020)	EMNLP20	Extraction	Classification	✓	✓	✓	Average
(Li et al., 2020)	COLING20	Extraction	Classification	✓	✓	✓	Average
(Zhang et al., 2020)	COLING20	Extraction	Classification	✓	✓	✓	Average
(Zhou et al., 2020)	COLING20	Extraction	Classification	✓	✓	✓	Average
(Christopoulou et al., 2019)	EMNLP19	Extraction	Classification	✓	✓	✓	Average
(Jia et al., 2019)	NAACL19	Extraction	Classification	✓	✓	✗	LogSumExp

Table 1: Recent DocRE models in anti-chronological order. "**Localization**", "**Linking**" and "**Typing**" indicates that an approach needs accurate annotations of entity localization, entity linking and entity typing, respectively. "**Aggregation**" indicates the strategy that how to aggregate multiple mention representations of an entity.

tillation (Tan et al., 2022) into transformer-based neural models. In addition, some studies (Soares et al., 2019; Zhou et al., 2021; Zhong and Chen, 2021; Zhou and Chen, 2022; Zhang et al., 2022) already verified that inserting special symbols (*e.g.,* [entity] and [/entity]) before and after named entities can significantly benefit relation representation encoding.

Observations from Literature Review. From Table 1, we have following key observations: (1) The listed studies claim that they address the problem of "document-level relation **extraction**"[2], but the relation **classification** is actually performed. (2) All graph-based approaches build homogeneous or heterogeneous graphs based on the unrealistic **precondition** that accurate annotations of entity localization, entity linking and entity typing are available. (3) Some pooling strategies (*e.g.,* Max, Average and LogSumExp) are widely used in modeling DocRE when aggregating representations of multiple mentions of an entity. However, it is unclear how the wrongly-detected mentions affect the performance of DocRE models.

3 Check on Dataset Annotations

To provide in-depth observations of the data assumption in most of DocRE models, we first take a thorough examination of data annotations on three commonly-used DocRE datasets. We will conduct quantitative and qualitative studies to analyze entity mentions and entity aliases which a relation instance involved.[3]

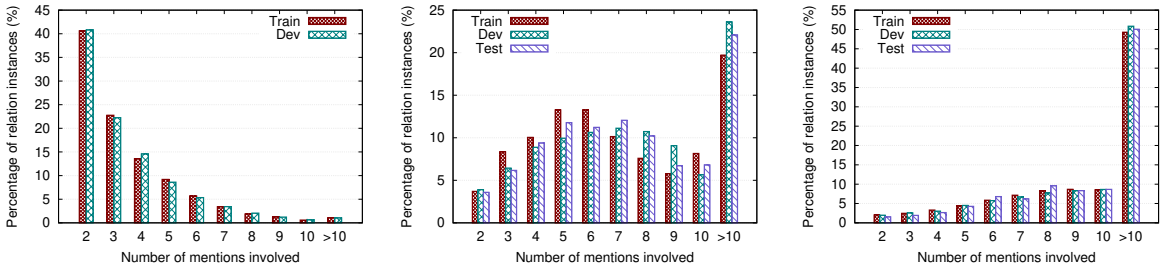
3.1 Probing Datasets

The summary of datasets is shown in Table 2. NA-instance means that there is no relation between head and tail entities. Non-NA instance means that there is at least one relation between head and tail entities. Note that the mention statistics in this Section are based on Non-NA instances.

DocRED (Yao et al., 2019) is a human-annotated dataset from Wikipedia and Wikidata. DocRED has 5,053 documents, 97 relation classes, 132,275

[2]The term "Extraction" commonly refers to extract relation types, head and tail entities from raw text.

[3]Entity Mentions: The words in text that refer to an entity. Entity Aliases: Unique mentions of an entity. Relation Instance: A piece of text involving head and tail entities to be classified.



(a) DocRED (b) CDR (c) GDA

Figure 2: Entity mention statistics on three datasets.

Datasets		#Doc.	#Rel.	#Non-NA
DocRED	Train	3,053	97	385,272
	Dev	1,000	97	11,518
	Test	1,000	97	-
CDR	Train	500	2	1,055
	Dev	500	2	1,025
	Test	500	2	1,087
GDA	Train	23,353	2	36,079
	Dev	5,839	2	8,762
	Test	1,000	2	1,502

Table 2: Statistics of DocRE datasets. (#Doc.: number of documents, #Rel.: number of relation labels, #Non-NA: number of non-NA-relation instances.)

entities, and 56,354 relational facts in total. The average length of documents in DocRED is around 8 sentences. Following previous studies (Yao et al., 2019; Wang et al., 2019), we use the standard split of the dataset: 3,053 documents for training, 1,000 for development and 1,000 for test.

CDR (Li et al., 2016) consists of three separate sets of articles with diseases, chemicals and their relations annotated. There are two relation labels: None and Chemical-Disease. There are total 1,500 articles and 500 each for the training, development and test sets.

GDA (Wu et al., 2019) is a Gene-Disease Association dataset from MEDLINE abstracts: 29,192 articles for training and 1,000 for testing. Following previous studies (Christopoulou et al., 2019; Li et al., 2021b), we further split the original training set into two sets: 23,353 for training and 5,839 for development. There are two relation labels: None and Gene-Disease.

3.2 Data Observations and Findings

We organize our findings by answering following research questions (RQs):

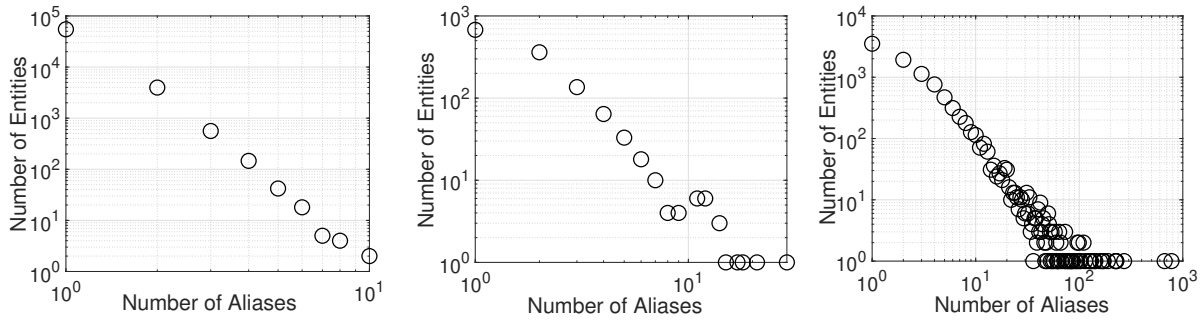
(RQ1): How many entity mentions are involved in a relation instance in commonly-used DocRE datasets?

We define that a relation instance to be classified is a piece of text containing head and tail entities. Thus, it is natural that the head or tail entity may have multiple mentions in the document. Figure 2(a), 2(b) and 2(c) show entity mention statistics in DocRED, CDR and GDA, respectively. The horizontal axis shows number of mentions of a relation instance. The vertical axis shows the percentages of relation instances in datasets. In the DocRED dataset, 59.2% of relation instances have more than two mentions. For CDR, 96% of relation instances have more than two mentions and 21% of relation instances have more than 10 mentions. For GDA, 98% of relation instances have more than two mentions and 50% of relation instances have more than 10 mentions. Our this finding reveals the huge difference between the sentence-level and document-level RE. That is, document-level RE involves much more entity mentions than sentence-level RE because of the longer text in document-level RE. One strong (almost untenable) assumption of existing DocRE models is that all entity mentions of a relation instance are successfully identified.

(RQ2): How many aliases does an entity have in commonly-used DocRE datasets?

RQ1 already showed that a relation instance may have multiple entity mentions. A follow-up question is about the number of unique mentions. Given that an entity can appear multiple times in a document, we define that the aliases of an entity are unique mentions. We are interesting in how many aliases an entity has.

Figure 3 plots the distribution of number of entities to number of aliases on three commonly-used datasets. For DocRED, we can observe that most of entities have only one alias and 4,745 entities



(a) DocRED (b) CDR (c) GDA

Figure 3: Statistics of entity aliases on three datasets.

have more than one alias. The maximum number of aliases is 10. For CDR, 650 entities (account for 48.95%) have more than one alias and the maximum number of aliases is 29. For GDA, 5,927 entities (account for 62.73%) have more than one alias and the maximum number of aliases is 778. CDR and GDA have more diverse aliases than DocRED, because DocRED is constructed from Wikipedia while CDR and GDA are constructed from biomedical text. Linking diverse aliases of an entity to its identifier is a challenging task in a long document. Our findings identify the strong (almost untenable) assumption of existing DocRE models that all the aliases (*i.e.,* unique mentions) of an entity are successfully normalized (*i.e.,* linked to its unique identifier).

(RQ3): Do the aliases of an entity vary widely in commonly-used DocRE datasets?

RQ2 already showed that an entity may have multiple aliases. For example, an entity in GDA has 778 unique aliases. In this Section, we investigate whether the aliases of an entity vary widely. Table 3 shows details of entity aliases ranked by numbers of aliases in the three datasets. For DocRED, the variation of aliases is slight because the genre of text is from formal articles. Although DocRED is manually annotated by human beings, there are still some annotation errors on entity linking. As shown in Table 3, the entity (Q180611, Azpeitia) is linked to many wrong aliases such as "United States" and "Chile". This observation confirms that document-level entity linking is a very challenging task. For biomedical datasets, chemical entities have a slight variation of aliases while gene and disease entities have a huge variation of aliases. In addition, there are many abbreviations for biomedical entities. Thus, effective NER and entity linking are key preconditions in modeling DocRE. We will further investigate the model usability with NER and entity linking in Section § 5.

4 Check on Model Robustness

Most of existing DocRE models are proposed based on the strong assumptions of mention annotations as shown in Section § 3. In this Section, we are interested in the following research question:

(RQ4): Are neural DocRE models robust to entity mention attacks?

To answer **RQ4**, we adopt **behavioral probing** (Lasri et al., 2022; Chen et al., 2022) to observe a model's behaviors by studying the model's predictions on attacking datasets. That is, attacks are only added at test time and are not available during model training.

4.1 Attacking Target Models

We investigate three typical DocRE models: (1) **BiLSTM-Sum** (Yao et al., 2019) which uses BiLSTM to encode the document and computes the representation of an entity by summing the representations of all mentions. (2) **GAIN-Glove** (Zeng et al., 2020) which constructs a heterogeneous mention-level graph and an entity-level graph to capture document-aware features and uses GloVe (Pennington et al., 2014) as word embeddings. (3) **BERT-Marker** (Zhou et al., 2021; Zhou and Chen, 2022) which takes BERT as the encoder and inserts special entity symbols before and after entities. More details of attacking target models can be found in Appx. § B.1.

4.2 Attack Construction

In this work, we focus on entity mention attacks which add data perturbations by taking into ac-

Rank	IDs	#Aliases	Details of Entity Aliases
DocRED Dataset			
1	Q180611 (LOC)	10	Azpeitia, Guipuzcoa, Cuba, Mexico, Azkoitia, Basque Country, United States, Argentina, Chile, Spain
2	Q544565 (LOC)	10	Qu, Yuxi River, Jiuxi River, Zhuji River, Ni River, Eshan River, Liucun River, Huaxi River, Qu River, Zhou River
3	Q3738980 (LOC)	8	Toding, Tsanda, Tsada, Tholing, Zanda, Toling, Zada, Tuolin
4	Q12274473 (MISC)	8	wazīrwāla, Waziri, Maseedwola, Wazirwola, Dawarwola, Wazir, of the Wazirs, Waziri Pashto
CDR Dataset			
1	D016572 (Chemical)	7	cyclosporine, cyclosporin, CsA, Cyclosporine, CyA, cyclosporin A, cyclosporine A
2	D014635 (Chemical)	7	divalproex sodium, VPA, sodium valproate, Valproic acid, Valproate, valproic acid, valproate
1	D007674 (Disease)	29	renal damage, CAN, nephrotoxic, renal dysfunctio, renal injury, Nephrotoxicity, kidney diseases, liver or kidney disease, cardiac and renal lesions, glomerular injury, kidney damage, ...omit...
2	D056486 (Disease)	21	Hepatitis, drug-induced hepatitis, acute hepatitis-like illness, liver damage, hepatotoxicity, cholestatic hepatitis, hepatic damage, hepatocellular injury, Toxic hepatitis, Granulomatous hepatitis, ...omit...
GDA Dataset			
1	348 (Gene)	114	apolipoprotein e4, APOE*4, ApoE2, apolipoprotein gene E4 allele, ApoE-4, apoE 4, apolipoprotein-E gene, Apolipoprotein E-epsilon4, factor–apolipoprotein E, Apolipoprotein (apo)E, ...omit...
2	7124 (Gene)	83	tumor necrosis factor alpha, tumor necrosis factor beta, Interleukin-1 and tumor necrosis factor-alpha, tumor necrosis factor alpha, TNF-)a, Tumor Necrosis Factor, TNF-308G/A, miR-21, IL6 and TNF, ...omit...
1	D030342 (Disease)	778	inherited defect of fatty acid oxidation, genetic haemochromatosis, inherited skin disorders, A-related disorders, autosomal-recessive pleiotropic disorder, autosomal dominant juvenile ALS, ...omit...
2	D009369 (Disease)	668	mammary tumors, tumor suppressor genes, MSI-H cancers, rectal cancers, predominant in lung tumour, early-stage prostate cancer, Tumour-necrosis, Cervical cancer, Malignant tumors, distal tumors, ...omit...

Table 3: Details of entity aliases ranked by number of aliases.

count different types of wrongly-detected mentions. The ultimate goal is to test the model robustness under different mention attacks. Therefore, we construct four types of attacks: (1) `DrpAtt`: we simply drop 50% of mentions of an entity if the entity has more than one mention. This attack is designed to simulate the case of missed detections in NER systems. (2) `BryAtt`: we slightly move the ground boundaries of 50% of mentions of an entity if the entity has more than one mention (*e.g.,* "$[\![$Spanish Civil War$]\!]_{\text{MISC}}$ in" is changed to "Spanish $[\![$Civil War$]\!]_{\text{MISC}}$ in"). (3) `CorAtt`: we intentionally make the coreference (*i.e.,* entity linking) of an entity wrong (*i.e.,* 50% of mentions of an entity are wrongly coreferential if the entity has more than one mention). (4) `MixAtt`: this attack is the mix of aforementioned three attacks. More attack details can be found in Appx. § B.2.

4.3 Attacking Results and Analysis

Table 4 reports the performance on various entity mention attacks for three attacking target models. We have the following observations:

First, all target models are significantly affected by the four attacks, with relative F1 drops from 7.93% to 85.51%. Overall, GAIN-Glove and BERT-Marker are more vulnerable than BiLSTM-Sum. This is because BERT-Marker requires accurate mention positions for inserting entity markers and GAIN-Glove needs the information of mention positions and normalization for constructing heterogeneous graphs. More specifically, BERT-Marker averagely suffers drops of 44.42%, 71.58%, and 71.72% across all attacks on DocRED, CDR and GDA, respectively. BiLSTM-Sum averagely suffers drops of 23.22%, 35.40%, and 40.67% across all attacks on DocRED, CDR and GDA, respectively.

Second, the `MixAtt` attack leads to more significant drops in performance for all attacking target models. `CorAtt` is more significant to impact robustness than `BryAtt` and `DrpAtt`. For instance, `CorAtt` leads to relative drops of 40.81%, 56.76% and 64.48% across three target models on DocRED,

Model	Attack	DocRED		CDR		GDA	
		F1%	Δ%	F1%	Δ%	F1%	Δ%
BiLSTM-Sum	No Attack	49.32	-	53.67	-	75.87	-
	DrpAtt	42.55	-13.73	48.34	-9.93	66.55	-12.28
	BryAtt	39.04	-20.84	39.23	-26.91	57.30	-24.48
	CorAtt	37.21	-24.55	28.86	-46.23	31.32	-58.72
	MixAtt	32.67	-33.76	22.26	-58.52	24.88	-67.21
GAIN-Glove	No Attack	54.91	-	55.13	-	78.65	-
	DrpAtt	48.17	-12.27	50.76	-7.93	59.61	-24.21
	BryAtt	41.82	-23.84	36.33	-34.10	45.92	-41.61
	CorAtt	32.34	-41.11	27.40	-50.30	33.24	-57.74
	MixAtt	28.56	-47.99	18.34	-66.73	23.52	-70.10
BERT-Marker	No Attack	59.82	-	64.47	-	82.71	-
	DrpAtt	47.34	-20.86	25.57	-60.34	36.43	-55.95
	BryAtt	41.45	-30.71	21.46	-66.71	24.55	-70.32
	CorAtt	25.86	-56.77	16.93	-73.74	19.04	-76.98
	MixAtt	18.34	-69.34	9.34	-85.51	13.55	-83.62

Table 4: Results of mention attacks on three datasets. Δ% indicates the relative performance changes between mention attacks and the original input ("No Attack").

CDR and GDA, respectively. DrpAtt leads to relative drops of 15.62%, 26.07% and 30.81% across three target models on DocRED, CDR and GDA, respectively. Our empirical results clearly show that the information of entity coreference, boundary and position plays an important role in DocRE.

Overall, based on the robustness evaluation in Table 4, we can answer **RQ4**: Most of neural DocRE models are far away from robustness to entity mention attacks. Therefore, it has some realistic significance to challenge current problem setups regarding data annotation assumptions in DocRE and to improve the robustness of DocRE models on entity mention attacks.

5 Check on Model Usability

In this Section, we investigate this realistic situation: DocRE models are already trained and training data is unavailable. We want to extract same relations on unseen raw text using these models. The goal is to deploy the already-trained DocRE models in other NLP applications. Here, we are interested in the following research question:

(RQ5): Are existing DocRE models easily adopted in real-world DocRE scenarios?

To answer **RQ5**, a necessary step is that whether we can process the raw text with the format as DocRE models trained on. This preprocessing procedure involves two crucial systems: Named Entity Recognition (NER) and Entity Linking.

5.1 Check on NER

Setups. Assume that DocRE models are already trained and the training sets are unavailable. We take the raw text of development set of DocRED, and test sets of CDR and GDA as the unseen data. We use strict match metrics (*i.e.,* entity boundary and type are both correctly detected) to measure agreement between the annotations we preprocessed and existing ground truth annotations.

NER Systems. For DocRED, we adopt three off-the-shelf NER systems: Flair (Akbik et al., 2019) and spaCy[4] and Stanza (Qi et al., 2020). For CDR and GDA, we adopt three biomedical NER systems: HunFlair (Weber et al., 2021), Stanza biomedical models (Zhang et al., 2021c) and Scispacy (Neumann et al., 2019). More details of NER systems can be found in Appx. § B.3.

Results on NER. Table 5 reports experimental results of NER systems on the three datasets. For DocRED, Flair achieves the best performance by the F1 score of 63.47%. Although the genre of DocRED is the formal text (*i.e.,* Wikipedia), the state-of-the-art NER systems are still unable to achieve decent performance on DocRED. HunFlair gets the best performance on the biomedical datasets because it trained on harmonized versions of 31 biomedical datasets.

[4] https://spacy.io/

Dataset	NER System	Strict Match (%) P	R	F1
DocRED	Flair	62.88	64.07	63.47
	spaCy	62.86	59.58	61.17
	Stanza	56.96	58.44	57.69
CDR	HunFlair	94.59	94.14	94.36
	Stanza	86.80	87.94	87.37
	ScispaCy	84.93	80.32	82.56
GDA	HunFlair	79.11	84.74	81.83
	Stanza	69.87	79.70	74.47
	ScispaCy	68.61	64.61	66.55

Table 5: Results of NER systems.

Dataset	Linking System	Strict Match (%) P	R	F1
DocRED	TagMe, ρ=0.1	24.2	42.5	30.8
	TagMe, ρ=0.2	35.0	38.6	36.7
	TagMe, ρ=0.3	45.7	33.5	38.7
	TagMe, ρ=0.4	52.4	27.8	36.4
	TagMe, ρ=0.5	49.7	12.4	19.8
CDR	ScispaCy, mesh	42.4	60.6	49.9
	ScispaCy, umls	53.7	63.3	58.1
GDA	ScispaCy, mesh	31.5	28.4	29.8
	ScispaCy, umls	30.9	38.6	34.3

Table 6: Results of entity linking systems. ρ is the confidence score (annotations that are below the threshold will be discarded). "mesh" and "umls" mean that entities are linked to the Medical Subject Headings and the Unified Medical Language System, respectively.

5.2 Check on Entity Linking

Setups. We examine the capability of entity linking systems on reproducing ground truth annotations for development/test sets of DocRED, CDR and GDA. We choose the strict match as the metric that a linking prediction is regarded as correct only if all mentions of an entity are correctly linked to the entity.

Entity Linking Systems. Unlike NER systems, there are very few off-the-shelf linking systems available. We choose TagMe (Ferragina and Scaiella, 2010) as the linker for DocRED, and Scispacy (Neumann et al., 2019) for CDR and GDA. More details of entity linking systems can be found in Appx. § B.4.

Results on Entity Linking. Table 6 reports experimental results of entity linking systems on the three datasets. For TagMe, the precision increases gradually with the increase of the value of ρ (confidence score), while the recall decreases as ρ increases. The best F1 on DocRED is only 38.7% with a confidence score of 0.3. Scispacy achieves F1 scores of 58.1% and 34.3% using umls for CDR and GDA, respectively. One key observation drawn from Table 6 is that document-level entity linking is a challenging task and existing linking systems commonly perform poorly on this task.

Based on empirical results of Sections 5.1 and 5.2, we can answer **RQ5**: Most of existing DocRE models are difficult to be adopted in real-world DocRE scenarios due to the need of input preparation for each pipeline module and the accumulation of errors in NER and entity linking systems.

6 Discussion

Let's Stop Simplifying Problem Setups. As summarized in Table 1, recent advances from the past four years have claimed significant progress in DocRE performance. However, our study shows that the actual improvements are attributable to a strong or even untenable assumption where all entities are perfectly typed, localized and normalized. Therefore, high F1 scores on leaderboards do not mean that the task of DocRE has been solved. Based on our findings (§4 and §5), the simplified problem setups cannot cover realistic scenarios. Even worse, the problem simplification significantly hurts the usability of deploying DocRE models in real-world end-user NLP applications. We call attentions on the community to address the real DocRE problem under the open-world assumption, rather than to push up the boundaries of simplified benchmarks for leading leaderboards.

Let's Model DocRE in the Wild. As shown in Section § 5, it is very difficult to produce accurate data formats as existing DocRE models trained on. Thus, given a new document, we are still unable to easily deploy existing trained DocRE models to extract same types of relations, let alone unseen relations. Recently, some studies (Cabot and Navigli, 2021; Eberts and Ulges, 2021; Giorgi et al., 2022) have started exploring the direction of jointly extracting entities and relations at document level. However, the end-to-end performance at document level is much worse than the performance at sentence level. Our empirical findings call more atten-

tions on developing high-performance end-to-end DocRE models and more attentions on modeling DocRE in the wild, rather than in an unrealistic Utopian world.

7 Conclusion

In this paper, we try to answer whether the performance gains recent DocRE models claimed are actually true. We took a comprehensive literature review of DocRE models and a thorough examination of popular DocRE datasets. We investigated the model robustness under four types of mention attacks and the model usability under a more realistic setting. Our findings call future efforts on modeling DocRE in the wild.

Limitations

We have discussed the implications of our research in Section 6. In this Section, we further discuss the threats to validity of our study.

- **Threats to Internal Validity**: The main internal threat to the validity of our research comes from **(RQ3)** where we present a qualitative study on the variation of aliases. We are unable to cover all cases in the qualitative study. For example, the entity of D030342 (Disease) in Table 3 has 778 unique aliases. It is impossible to show all aliases to readers. To help mitigate this threat, we try to show as many examples as possible in a limited space.
- **Threats to External Validity**: The main threat to external validity arises from the potential bias in the selection of experimental datasets, attacking target models and off-the-shelf NER and Entity Linking tools. To mitigate this threat, we experiment with multiple datasets, models and tools. For experimental datasets, we choose the three most popular DocRE datasets (*i.e.,* DocRED, CDR, and GDA). We believe that these three datasets are broadly representative in this research community. For attacking target models, we choose three typical models ranging from non-contextualized sequence-based to graph-based, and to contextualized Transformers models. For off-the-shelf NER/Linking tools, we comprehensively investigate five state-of-the-art NER taggers and two entity linkers.

Ethical Considerations

As our goal of this study is to challenge current problem setups of DocRE, we heavily rely upon existing well-known datasets, models and NLP tools. We only claim that our findings may hold on similar datasets or domains. We acknowledge the risk of generalizability of our findings on other privacy-sensitive datasets or specific domains. In general, we suggest that practitioners repeat all experiments following our procedures when using other corpora.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT demo)*, pages 54–59.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: relation extraction by end-to-end language generation. In *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2370–2381.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? In *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3792–3805.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4924–4935.

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

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *The 19th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1625–1628.

John Giorgi, Gary D. Bader, and Bo Wang. 2022. A sequence-to-sequence approach for document-level relation extraction. In *The 21st Workshop on Biomedical Language Processing, (BioNLP)*, pages 10–25.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021. Three sentences are all you need: Local path enhanced docu-

ment relation extraction. In *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 998–1004.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3693–3704.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the usage of grammatical number. In *The 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8818–8831.

Bo Li, Wei Ye, Canming Huang, and Shikun Zhang. 2021a. Multi-view inference for relation extraction with uncertain knowledge. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 13234–13242.

Bo Li, Wei Ye, Zhonghao Sheng, Rui Xie, Xiangyu Xi, and Shikun Zhang. 2020. Graph enhanced dual attention network for document-level relation extraction. In *The 28th International Conference on Computational Linguistics (COLING)*, pages 1551–1560.

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 J. Biol. Databases Curation*.

Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. 2021b. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 1359–1370.

Kohei Makino, Makoto Miwa, and Yutaka Sasaki. 2021. A neural edge-editing approach for document-level relation graph extraction. In *Findings of The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 2653–2662.

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

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *The 18th BioNLP Workshop and Shared Task (BioNLP@ACL)*, pages 319–327.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *The 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL demo)*, pages 101–108.

Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. 2021. Learning logic rules for document-level relation extraction. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1239–1250.

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

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1672–1681.

Hieu Minh Tran, Trung Minh Nguyen, and Thien Huu Nguyen. 2020. The dots have their values: Exploiting the node-edge connections in graph-based neural models for document-level relation extraction. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4561–4567.

Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *The 57th Conference of the Association for Computational Linguistics (ACL)*, pages 229–240.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *The 6th International Conference on Learning Representations (ICLR)*.

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

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

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. Hunflair: an easy-to-use tool for state-of-the-art

biomedical named entity recognition. *Bioinform.*, 37(17):2792–2794.

Ye Wu, Ruibang Luo, Henry C. M. Leung, Hing-Fung Ting, and Tak Wah Lam. 2019. RENET: A deep learning approach for extracting gene-disease associations from literature. In *The 23rd Annual International Conference Research in Computational Molecular Biology (RECOMB)*, volume 11467, pages 272–284.

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

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 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 257–268.

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

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on freebase via relation extraction and textual evidence. In *The 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021b. Discriminative reasoning for document-level relation extraction. In *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 1653–1663.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021c. Document-level relation extraction with reconstruction. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 14167–14175.

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 *The 57th Conference of the Association for Computational Linguistics (ACL)*, pages 764–777.

Jiaxin Yu, Deqing Yang, and Shuyu Tian. 2022. Relation-specific attentions over entity mentions for enhanced document-level relation extraction. In *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. SIRE: separate intra- and inter-sentential reasoning for document-level relation extraction. In *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 524–534.

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

Liang Zhang, Jinsong Su, Yidong Chen, Zhongjian Miao, Zijun Min, Qingguo Hu, and Xiaodong Shi. 2022. Towards better document-level relation extraction via iterative inference. In *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 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 *The Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3999–4006.

Sheng Zhang, Cliff Wong, Naoto Usuyama, Sarthak Jain, Tristan Naumann, and Hoifung Poon. 2021b. Modular self-supervision for document-level relation extraction. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5291–5302.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. 2021c. Biomedical and clinical english model packages in the stanza python NLP library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.

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

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 50–61.

Huiwei Zhou, Yibin Xu, Weihong Yao, Zhe Liu, Chengkun Lang, and Haibin Jiang. 2020. Global context-enhanced graph convolutional networks for

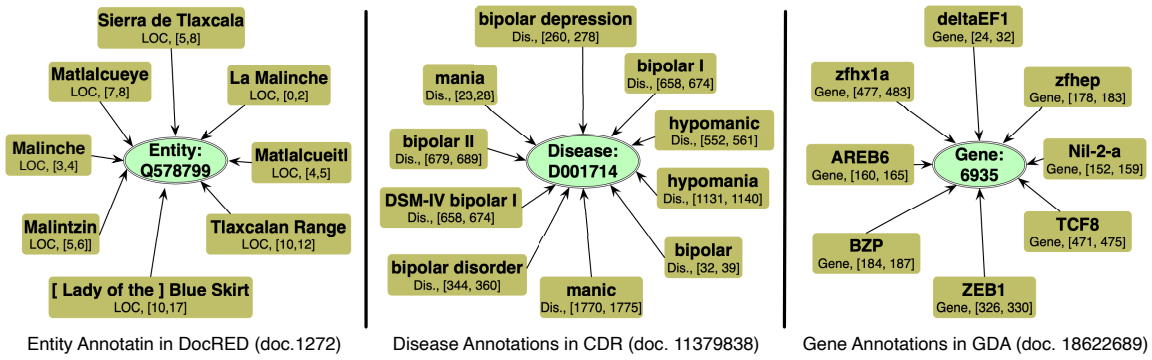


Figure 4: Additional examples of data assumption in three popular DocRE datasets. Entities are annotated with types (LOC/Disease/Gene), positions ([start, end]) and unique identifiers (Q578799/D001714/6935).

document-level relation extraction. In *The 28th International Conference on Computational Linguistics (COLING)*, pages 5259–5270.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, pages 161–168.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 14612–14620.

Appendix

A Additional Examples of Data Annotations

Our study identifies a strong or even untenable assumption in DocRE. To give more intuitive sense, Figure 4 shows additional examples of data assumption in three popular DocRE datasets. Specifically, the eight entity mentions (*i.e., Tlaxcalan Range, Matlalcueitl, [Lady of the] Blue Skirt, Malintzin, Sierra de Tlaxcala, Malinche, La Malinche, Matlalcueye*) are annotated with types and positions, then linked to a unique identifier in the DocRED corpus. The ten entity mentions (*mania, bipolar II, bipolar I, bipolar depression, hypomanic, hypomania, DSM-IV bipolar I, bipolar, manic, bipolar disorder*) are typed, localized and normalized in the CDR corpus. The eight entity mentions (*deltaEF1, zfhx1a, zfhep, AREB6, Nil-2-a, BZP, TCF8, ZEB1*) are typed, localized and normalized in the GDA corpus. Most of existing DocRE models are developed based on the assumption that all entity mentions are perfectly typed, localized and normalized.

B More Experimental Details

B.1 Attacking Target Models

BiLSTM-Sum. BiLSTM-Sum (Yao et al., 2019) uses a bidirectional LSTM to encode documents and computes the representation of an entity by summing the representations of all mentions. The embeddings from `glove.840B.300d`[5] are used to initialize model vocabularies for DocRED, CDR and GDA. All word embeddings and model parameters are learnable during training. Hyperparameters are tuned on the development set for each dataset respectively.

GAIN-Glove. GAIN-Glove (Zeng et al., 2020) constructs a heterogeneous mention-level graph to model complex interaction among different mentions across the document. Then a path reasoning mechanism is proposed to infer relations between entities based on another constructed entity-level graph. We implement GAIN-Glove with 2 layers of GCN and the dropout rate of 0.6 based on the codes[6]. The embeddings from `glove.840B.300d`[7] are used for DocRED, CDR and GDA.

BERT-Marker. BERT-Marker (Zhou et al., 2021; Zhong and Chen, 2021; Zhou and Chen, 2022) first inserts special entity symbols (*i.e.,* `[ent]` and `[/ent]`) before and after entities, then encodes the whole document using the pretrained BERT. The representation of token `[CLS]` is used for classification. In particular, we use the checkpoint `bert-base-uncased`[8]

[5] https://nlp.stanford.edu/projects/glove/
[6] https://github.com/DreamInvoker/GAIN
[7] https://nlp.stanford.edu/projects/glove/
[8] https://huggingface.co/

for DocRED, and the checkpoint `allenai/scibert_scivocab_uncased`[9] for CDR and GDA.

All attacking target models are implemented with `PyTorch`[10] and `Accelerate`[11], and trained on one DGX machine, totally equipped with 80 Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz processor cores, 400 GB of RAM, and 8 NVIDIA Tesla V100-32GB GPUs.

B.2 Attack Details

In total, we construct four types of attacks, *i.e.,*, `DrpAtt`, `BryAtt`, `CorAtt` and `MixAtt`, to check the robustness of attacking target models.

`DrpAtt`. Missing some entities is a very common phenomenon for most of NER systems. `DrpAtt` is constructed to investigate the effect of missed mentions. If an entity has more than one mention, we simply drop 50% of mentions of the entity.

`BryAtt`. Some entities are complex and nested in natural language. Detecting boundaries precisely is not a trivial task. `BryAtt` is constructed to investigate the effect of wrongly-detected entity boundaries. If an entity has more than one mention, we slightly move the ground boundaries of 50% of mentions of the entity.

`CorAtt`. The document-level coreference resolution is a challenging task in DocRE. Most of existing DocRE models are developed on benchmark datasets where entity coreference is manually annotated. `BryAtt` is constructed to investigate the effect of wrongly-coreferential mentions. We intentionally make the coreference results (*i.e.,* entity linking) of an entity wrong (*i.e.,* 50% of mentions of an entity are wrongly coreferential if the entity has more than one mention).

`MixAtt`. This type of attack is the mix of aforementioned three attacks.

B.3 NER Systems

In Section 5.1, we adopt five off-the-shelf NER systems in our experiments.

Flair. Flair[12] is a very simple framework for state-of-the-art NLP and developed by Humboldt University of Berlin and friends. We use the `ner-english-ontonotes-large`[13] model for DocRED.

spaCy. spaCy[14] is a library for advanced Natural Language Processing in Python and Cython. We use the `en_core_web_trf`[15] model for DocRED.

Stanza. Stanza[16] is a collection of accurate and efficient tools for the linguistic analysis of many human languages, developed by Stanford NLP Group. General domain, biomedical & clinical models are available in Stanza. We use the `ontonotes`[17] for DocRED, `bc5cdr`[18] for CDR, `bc5cdr` and `bionlp13cg` for GDA.

HunFlair. HunFlair[19] is a state-of-the-art NER tagger for biomedical texts. It contains harmonized versions of 31 biomedical NER datasets. We use `hunflair-chemical` and `hunflair-disease` for CDR, `hunflair-gene` and `hunflair-disease` for GDA.[20]

ScispaCy. ScispaCy[21] is a Python package containing spaCy models for processing biomedical, scientific or clinical text. We use `en_ner_bc5cdr_md` for CDR. We use `en_ner_bc5cdr_md`, and `en_ner_bionlp13cg_md` for GDA.[22]

B.4 Entity Linking Systems

Comparing with flourishing NER systems, there are very few entity linking systems available. We adopt two widely-used entity linking systems in our experiments.

TagMe. TagMe[23] is a powerful tool that identifies on-the-fly meaningful substrings (called "spots") in

`bert-base-uncased`

[9] https://huggingface.co/allenai/scibert_scivocab_uncased

[10] https://pytorch.org/

[11] https://github.com/huggingface/accelerate

[12] https://github.com/flairNLP/flair

[13] https://huggingface.co/flair/ner-english-ontonotes-large

[14] https://spacy.io/

[15] https://spacy.io/models/en#en_core_web_trf

[16] https://stanfordnlp.github.io/stanza/

[17] https://stanfordnlp.github.io/stanza/ner_models.html

[18] https://stanfordnlp.github.io/stanza/available_biomed_models.html

[19] https://github.com/flairNLP/flair/blob/master/resources/docs/HUNFLAIR.md

[20] https://github.com/flairNLP/flair/blob/master/flair/models/sequence_tagger_model.py#L751

[21] https://allenai.github.io/scispacy/

[22] https://github.com/allenai/scispacy

[23] https://sobigdata.d4science.org/web/tagme/tagme-help

an unstructured text and link each of them to a pertinent Wikipedia page in an efficient and effective way. We use the official Python TagMe API wrapper[24] for DocRED. The confidence scores (annotations that are below the threshold will be discarded) are experimented among $[0.1, 0.2, 0.3, 0.4, 0.5]$.

Entity Linker in ScispaCy. Entity Linker in ScispaCy[25] is a spaCy component which performs linking to a knowledge base. The linker simply performs a string overlap - based search (char-3grams) on named entities, comparing them with the concepts in a knowledge base using an approximate nearest neighbours search. For CDR and GDA datasets, we explore the following two knowledge bases:

- umls: Links to the Unified Medical Language System, levels 0,1,2 and 9. This has 3 million concepts.
- mesh: Links to the Medical Subject Headings. This contains a smaller set of higher quality entities, which are used for indexing in Pubmed. MeSH contains 30k entities.

C License

DocRED is released under The MIT license. GDA is released uner The GNU Affero General Public License. GAIN is released under The MIT License. Flair is released under The MIT License. spaCy is released under The MIT License. Stanza is Licensed under The Apache License 2.0. HunFlair is Licensed under The MIT License. ScispaCy is Licensed under The Apache License 2.0. TagMe is Licensed under The Apache License 2.0. PyTorch is with The Copyright (c) 2016 - Facebook, Inc (Adam Paszke). Huggingface Transformer models are released under The Apache License 2.0. All the scientific artifacts are consistent with their intended uses.

[24] https://github.com/marcocor/tagme-python

[25] https://github.com/allenai/scispacy#entitylinker

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations on Page 9.

☑ A2. Did you discuss any potential risks of your work?
Ethical Considerations on Page 9.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract on Page 1. 1 Introduction Section on Page 2.

☒ A4. Have you used AI writing assistants when working on this paper?
No AI writing assistant used.

B ☑ Did you use or create scientific artifacts?

Section 4.1 Section 5.1 Section 5.2

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1 Section 5.1 Section 5.2 Appx. B3 Appx. B4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appx. C

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appx. C

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Experiments are conducted on well-known benchmarks and follow previous studies.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.1 on Page 3.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Table 2 on Page 3.

C ☑ Did you run computational experiments?

Section 4.3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appx. B on Page 12.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appx. B on Page 12.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appx. B on Page 12.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appx. B on Page 12.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.