Did the Models Understand Documents? Benchmarking Models for Language Understanding in Document-Level Relation Extraction

Haotian Chen, Bingsheng Chen and **Xiangdong Zhou**
School of Computer Science, Fudan University
Shanghai Key Laboratory of Data Science
htchen18@fudan.edu.cn, bschen21@m.fudan.edu.cn, xdzhou@fudan.edu.cn

Abstract

Document-level relation extraction (DocRE) attracts more research interest recently. While models achieve consistent performance gains in DocRE, their underlying decision rules are still understudied: Do they make the right predictions according to rationales? In this paper, we take the first step toward answering this question and then introduce a new perspective on comprehensively evaluating a model. Specifically, we first conduct annotations to provide the rationales considered by humans in DocRE. Then, we conduct investigations and reveal the fact that: In contrast to humans, the representative state-of-the-art (SOTA) models in DocRE exhibit different decision rules. Through our proposed RE-specific attacks, we next demonstrate that the significant discrepancy in decision rules between models and humans severely damages the robustness of models and renders them inapplicable to real-world RE scenarios. After that, we introduce mean average precision (MAP) to evaluate the understanding and reasoning capabilities of models. According to the extensive experimental results, we finally appeal to future work to consider evaluating both performance and the understanding ability of models for the development of their applications. We make our annotations and code publicly available[1].

1 Introduction

Relation extraction (RE), aiming to extract relations between entities from texts, plays an important role in constructing a large-scale knowledge graph (Riedel et al., 2010; Hendrickx et al., 2010). Most previous work extract relations from a single sentence (Zelenko et al., 2002; Wei et al., 2020; Shang et al., 2022), while recent studies adopt multiple sentences as a whole to harvest more relations including inter-sentence relations (Yao et al., 2019), i.e., document-level relation extraction (DocRE).

[1] https://github.com/Hytn/DocRED-HWE

... Eva Bosáková and Věra Čáslavská led the **Czech** women' s gymnastics team to three successive World/Olympic silver medals in a row, establishing their nation as the foremost challengers to the dominant Soviet women's team during that era. ... she was World (1962) and Olympic (1960) champion, and she was good enough on all four events ...

Head:	**Věra Čáslavská**
Tail:	**Czech**
Prediction of Model:	*country of citizenship*
Ground Truth Relation:	*country of citizenship*
Evidence of Model:	and | Věra Čáslavská | 1960 | , | and
Evidence of Human:	their nation

Figure 1: An example from DocRED.

DocRE is more challenging because models are required to synthesize all information of a given document and then predict relations by reasoning and language understanding (Yao et al., 2019; Nan et al., 2020; Zeng et al., 2020).

Previous work obtains consistent performance gains on DocRED (Yao et al., 2019), the proposal of which has benefited the rapid development of DocRE (Huang et al., 2022). However, the extent to which their proposed methods possess language understanding and reasoning capabilities is still understudied. A common evaluation method is to measure average error across a test set, which neglects the situations where models can make right predictions according to wrong features. As shown in Figure 1, the model accurately predicts the relation between *Věra Čáslavská* and *Czech* as humans do. However, the evidence words considered by models are incomprehensible to humans. Similar situations, where models improve their performance by recognizing the spurious patterns, are identified by parts of the AI community, including *annotation artifacts* in natural language inference (NLI) (Poliak et al., 2018; Gururangan et al., 2018; Glockner et al., 2018) and *shallow template matches* in named entity recognition (NER) (Fu et al., 2020). These learned spurious patterns can severely damage their robustness and generalization abilities in the corresponding tasks (Geirhos

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6418–6435

et al., 2020). To the best of our knowledge, this is the first work to diagnose the decision rules of models in DocRE.

In this paper, we analyze and characterize the understanding ability of SOTA models in DocRE, expose the bottleneck of the models, and then introduce a new evaluation metric to select trustworthy and robust models from those well-performed ones. Our contributions are summarized as follows:

(1) We conduct careful and exhausting annotations on DocRED to propose $\text{DocRED}_{\text{HWE}}$, where HWE denotes human-annotated word-level evidence. The evidence words (decision rule) of humans are annotated in the dataset.

(2) We adopt a feature attribution method to observe the most crucial words considered by models in their reasoning processes. We reveal that the SOTA models spuriously correlate the irrelevant (non-causal) information (e.g., entity names, some fixed positions in any given documents, and irrelevant words) with their final predictions, forming their own unexplainable decision rules.

(3) We demonstrate that the decision rules of the SOTA models in DocRE are not reliable. We delicately design six kinds of RE-specific attacks to expose their bottleneck: Although they succeed in achieving improved performance on the held-out test set, they can strikingly fail under our designed attacks. Both the severe lack of understanding ability and the learned spurious correlations exacerbate the vulnerability of the models.

(4) Inspired by evaluation metrics in recommender systems, we evaluate the understanding and reasoning capability of models by our introduced mean average precision (MAP). MAP enables us to distinguish between the spurious-correlation-caused and the understanding-ability-caused improvements in the performance of models. We observe that a model with a higher MAP will achieve stronger robustness and generalization ability.

2 Related Work

Document-level Relation Extraction. Prevalent effective methods on document-level RE can be divided into two categories: graph-based methods and transformer-based methods (Huang et al., 2022). Both of them are based on deep neural networks (DNNs). Graph-based methods explore the structure information in context to construct various graphs and then model the process of multi-hop reasoning through the paths in graphs. According to the classification mentioned in previous work (Huang et al., 2022), the SOTA graph-based method is DocuNet (Zhang et al., 2021), which constructs an entity-level relation graph, and then leverages a U-shaped network over the graph to capture global interdependency. Transformer-based methods perform reasoning by implicitly recognizing the long-distance token dependencies via transformers. One of the most representative transformer-based methods is ATLOP (Zhou et al., 2020), which enhances the embeddings of entity pairs by relevant context and introduces a learnable threshold for multi-label classification. The techniques proposed by ATLOP are widely adopted by subsequent transformer-based work (Xie et al., 2022; Tan et al., 2022a; Xiao et al., 2022), including adaptive thresholding (AT) and localized context pooling (LOP).

Analyzing Decision Rules of DNNs. With the tremendous success and growing societal impact of DNNs, understanding and interpreting the behavior of DNNs has become an urgent necessity. In terms of NLP, While DNNs are reported as having achieved human-level performance in many tasks, including QA (Chen et al., 2019), sentence-level RE (Wang et al., 2020), and NLI (Devlin et al., 2018), their decision rules found by feature attribution (FA) methods are different from that of humans in many cases. For example, in argument detection, the widely adopted language model BERT succeeds in finding the most correct arguments only by detecting the presence of "not" (Niven and Kao, 2019). In VQA, dropping all words except "color" in each question is enough for a DNN to achieve 50% of its final accuracy (Mudrakarta et al., 2018). In NLI, DNNs can make the right predictions without access to the context (Poliak et al., 2018). It is demonstrated in these tasks that decision rules of models should approach that of humans. Otherwise, the difference will lead to a severe lack of robustness and generalization ability (Agrawal et al., 2016; Belinkov and Bisk, 2018; Fu et al., 2020). It remains understudied whether the same conclusion is established in DocRE. To the best of our knowledge, this is the first work comprehensively analyzing the decision rules of both models and humans in DocRE.

3 Data Collection

Our ultimate goal is to provide all of the evidence words (decision rules) that humans rely on

during the reasoning process in DocRE. Since it is not feasible for annotators to label relations and evidence from scratch in DocRE (Yao et al., 2019; Huang et al., 2022), we select DocRED to further annotate our fine-grained decision rule (word-level evidence). Our proposed dataset is named DocRED$_{\text{HWE}}$, where HWE denotes human-annotated word-level evidence. In the following two sections, we first elaborate on the underlying reasons why we conduct word-level evidence annotation and why on DocRED, and then introduce the details of our annotation.

3.1 Motivations

Motivation for Human Annotation. Current human annotations on DocRED are still insufficient to support our research: the evidence for each relational fact is sentence-level instead of word-level. If we base our study on the coarse-grained decision rules (sentence-level evidence) to analyze the reasoning behaviors of humans and models, the results will be misleading. For example, as shown in Figure 1, the sentence-level evidence of models and humans overlaps with each other (*and Věra Čáslavská* and *their nation* come from the same sentence), while their word-level evidence is totally different. Therefore, annotation of word-level evidence is of the essence. we conduct careful and exhausting word-level evidence annotation on DocRED and propose DocRED$_{\text{HWE}}$. Our proposed dataset significantly benefits more comprehensive analyses of DocRE, which will be discussed in Section 5.

Motivation for Selecting DocRED. While there are a few candidate datasets in DocRE, only one of them named DocRED (Yao et al., 2019) satisfies the urgent need of studying the understanding and reasoning capabilities of general-purpose models in real-world DocRE. Specifically, Quirk and Poon (2017) and Peng et al. (2017) leverage distant supervision to construct two datasets without human annotation, which hurts the reliability of the evaluation. Li et al. (2016) and Wu et al. (2019) proposed two human-annotated document-level RE datasets named CDR and GDA, respectively. Both of them serve specific domains and approaches (biomedical research) and contain merely one to two kinds of domain-specific relations. Different from other datasets in DocRE, the proposal of DocRED has significantly promoted the rapid development of the task in the past two years (Huang et al., 2022). The large-scale human-annotated dataset is constructed from Wikipedia and Wikidata, which serves general-purpose and real-world DocRE applications (Yao et al., 2019). Among various improved versions of DocRED (Huang et al., 2022; Tan et al., 2022b), we select the original version with annotation noise because it presents one of the most general circumstances faced by RE practitioners: having limited access to entirely accurate human-annotated data due to the extremely large annotation burden and difficulty. For example, human-annotated DocRED and TACRED (Zhang et al., 2017) are discovered to have labeling noise. As to distantly supervised datasets NYT (Mintz et al., 2009) and DocRED-distant, the amount of noise becomes larger.

3.2 Human Annotation Generation

Challenges and Solutions. We randomly sample 718 documents from the validation set of DocRED. Annotators are required to annotate all the words they rely on when reasoning the target relations. Note that we annotate the pronouns that can be another kind of mentions for each entity, which are crucial for logical reasoning but neglected in DocRED. Our annotation faces two main challenges. *The first challenge* comes from the annotation artifacts in the original dataset: Annotators can use prior knowledge to label the relations through entity names, without observing the context. For example, given a document with a cross-sentence entity pair "Obama" and "the US", annotators tend to label "president of" despite the lack of rationales. The issue is naturally solved by annotating the fine-grained word-level evidence. Consequently, despite the intensive workload, we annotate the words in reasoning paths for each relation. *The second challenge* lies in multiple reasoning paths for a single relation: Annotators are required to annotate the words in all reasoning paths. While annotators succeed in reasoning a certain relation through the corresponding evidence words, those words in other reasoning paths can often be neglected. To solve the issue, we adopt multiple (rolling) annotations for each document and propose the checking rule: Given a document and the previously annotated relation with its evidence words masked, the annotator will not be able to reason the relation. If the rule is violated, new evidence words will be annotated. The update will be checked by the next annotator until no update occurs. All of the

annotated evidence words are verified at least two times.

Quality of Annotation. To ensure the quality of the dataset, we provide principle guidelines and training to the annotators. We examine the annotators if they understand the principle. Meanwhile, we regularly inspect the quality of annotations produced by each annotator. Our inspection exerts a positive effect on the quality. On one hand, we filter out 18 out of 718 documents that present low annotation accuracy. Through the rolling annotation strategy, annotators also inspect the annotations from each other. On the other hand, annotators correct three kinds of annotation errors in the original DocRED: 1) relation type error where annotators wrongly annotate a relation type between an entity pair, 2) insufficient evidence error where an annotated relation can not be inferred from the corresponding document, and 3) evidence error where the sentence-level evidence of a relation is wrongly annotated. The number of errors in the three categories is 4, 44, and 90, respectively. We exhibit more details in the appendix.

4 Task, Methods, and Datasets

4.1 Task Description

Given a document d and an entity set $\mathcal{E} = \{e_i\}_{i=1}^{n}$ in d, the target of document-level relation extraction is to predict all of the relations between entity pair $(e_i, e_j)_{i,j=1...n;i\neq j}$ among $\mathcal{R} \cup \{\mathrm{NA}\}$. $\mathcal{R}$ is the predefined relations set . NA indicates that there is no relation between an entity pair. e_i and e_j denote subject and object entities. An entity may appear many times in a document, we use set $\left\{m_j^i\right\}_{j=1}^{N_i}$ to distinguish the mentions of each entity. We finally build the extracted relation triples into the form of $\{(e_i, r_{ij}, e_j) \mid e_i, e_j \in \mathcal{E}, r_{ij} \in \mathcal{R}\}$.

4.2 Methods

We choose one of the most representative models from each category of document-level RE models (DocuNet from graph-based methods and ATLOP from transformer-based methods) to produce attributions by feature attribution (FA) methods. We choose Integrated Gradient (IG) as our attribution method due to its verified simplicity and faithfulness (Sundararajan et al., 2017), which renders IG applicable in other text-related tasks (Mudrakarta et al., 2018; Liu and Avci, 2019; Bastings and Filippova, 2020; Hao et al., 2021; Liu et al., 2022).

Integrated Gradient Integrated Gradient is a reference-based method that calculates both the model output on the input and that on a reference point. The difference between the outputs is distributed as an importance score for each token. Specifically, given an input x and reference point x', IG computes the linear integral of the gradients g_i along the i^{th} dimension from x' to x by,

$$g_i = \left(x_i - x_i'\right) \times \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha \times (x - x')\right)}{\partial x_i} d\alpha,$$

where $\frac{\partial F(x)}{\partial x_i}$ indicates the gradient of an output $F(x)$ to x. As set in other text-related tasks (Wallace et al., 2019), we set x' as a sequence of embedding vectors with all zero values.

4.3 Datasets

DocRED and DocRED$_{\text{Scratch}}$. DocRED contains 56,354 human-annotated relational facts, which can be categorized into 96 relation types. Most of the relational facts (61.1%) can only be identified by reasoning (Yao et al., 2019). Recently, Huang et al. (2022) argue that the recommend-revise scheme adopted by DocRED in annotation leads to an obvious bias toward popular entities and relations. They rectify the bias by re-annotating 96 randomly selected documents (from the validation set of DocRED) from scratch and propose DocRED$_{\text{Scratch}}$. The distribution of DocRED$_{\text{Scratch}}$ shifts largely from the training set of DocRED, which renders it applicable for testing the generalization ability of models trained on DocRED.

DocRED$_{\text{HWE}}$ We propose DocRED$_{\text{HWE}}$ with the following features: 1) DocRED$_{\text{HWE}}$ contains 699 documents with 27,732 evidence words (10,780 evidence phrases) annotated by humans for 7,342 relational facts among 13,716 entities. 2) We annotate 1,521 pronouns referring to different entities, which are necessary to predict corresponding relations between entity pairs and neglected in DocRED. 3) At least 3,308 out of 7,342 (45.1%) relational facts require reading multiple sentences for extraction.

5 Experiment and Analysis

5.1 Analyzing Decision Rules of Models

We employ IG as our attribution technique to characterize the decision rules of models, which help us observe some potential risks in the SOTA models.

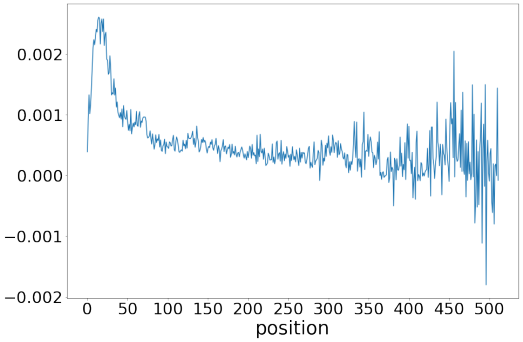

Figure 2: Mean attribution value distribution of $\text{ATLOP}_{\text{RoBERTa}}$ on different positions of documents in the validation set of DocRED. A similar shape of the curve emerges when attributing DocuNet. We only exhibit one of the curves due to the limited space.

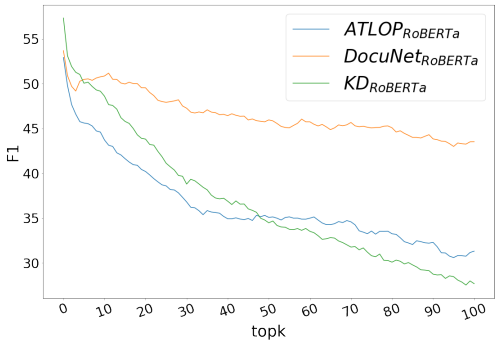

Figure 3: Performance based on top K attributed words

Position Discrimination. After being encoded by models, each token possesses its semantic meaning (word embedding) and position information (position embedding). Before analyzing the semantic meaning, we first visualize the contribution of position information to the predictions according to the attribution values. As shown in Figure 2, tokens in certain positions will affect final predictions more significantly than the words in other positions. In other words, models will discriminate words according to their positions in a document, even though the annotated rationales are almost uniformly distributed across the documents. We posit two reasons: (1) models distort the features from positions in the process of learning and spuriously correlate certain positions with the final predictions; (2) the position embeddings are wrongly trained (unsupervised), deviating from their original function of representing the position information. Furthermore, we observe more significant variances in those positions, roughly from 450 to 500, because the number of documents that are longer than 450 is small.

Note that the learned position discrimination may happen to apply to the test set of DocRED. However, the distributional shifts in real-world applications can render the spurious pattern no longer predictive. The generalization ability of models will be severely destroyed.

Narrow Scope of Reasoning. To observe the words that are necessary for a model to infer the right relations, we first investigate their number, representing the reasoning scope of models. Specifically, we design a template in the form of "A X B", where A and B denote the given entity pair and X can be either a word sequence or a single word. We regard X as necessary when models accurately predict the relation r_{AB} between A and B according to the template. We set X to the top K attributed tokens of r_{AB} and the position order of the tokens is the same as that in the original document. The performance of models on the validation set of DocRED is shown in Figure 3. Adding the highest attributed words surprisingly results in a performance decline. The contribution of position is significant, which is consistent with the results shown in Figure 2. Most importantly, we observe that models can achieve 53% F1-score when only given names of entity pairs without access to the context, which remains at about 85% of their original performance. Models perform reasoning in a strikingly narrow scope. If the phenomenon is reasonable, it indicates that such a few words are enough to explain rationales for the right predictions. To verify the assumption, We visualize these words in the next paragraph.

Spurious Correlations. We select the top five attributed words to visualize the evidence words of models shown in Figure 4. The attributions reveal that the SOTA models on DocRED largely rely on some non-causal tokens (e.g., entity name and some punctuations) to make the right predictions, which exerts a negative effect on learning the rationales. We can observe that the full-stop token, for example, plays a pivotal role in the predictions. Note that some special tokens ('[SEP]' and '[CLS]') are demonstrated to serve as "no-op" operators (Clark et al., 2019). The reliance on these special tokens may not be a problem because the two tokens are guaranteed to be present and are never attacked. However, the reliance on non-causal tokens

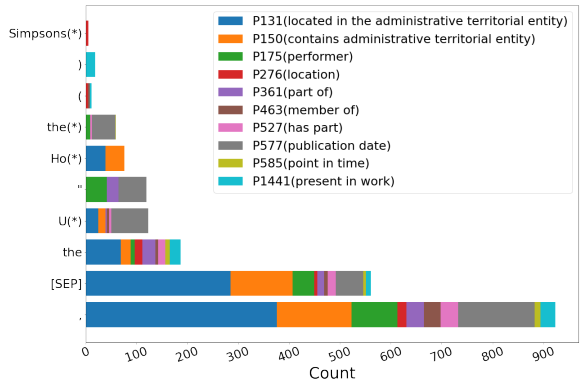

Figure 4: Statistics of word-level evidence of $\text{ATLOP}_{\text{RoBERTa}}$. The signal '*' denotes that the corresponding token belongs to entity names. We observe a similar phenomenon when counting the evidence words of DocuNet. We only exhibit one result due to the limited space.

renders a model fragile, untrustworthy, and far from being deployed into real-world scenarios as non-causal tokens can easily be attacked through substitutions, paraphrasing, changes in writing style, and so on. As shown in Table 1, if models learn to predict according to non-causal tokens, then each attack in these tokens will easily be successful. This severely destroys the robustness of models. The visualization indicates that models learn abundant spurious correlations (e.g., entity names and irrelevant words) to minimize the training error. We further prove that the spurious correlations are caused by selection bias in both pre-training and finetuning procedures. The details of the proof are given as follows.

Analysis of Underlying Causes. We shed some light on the underlying causes of learning spurious correlations. We argue that the common ground of the highly attributed non-causal tokens is that they are either high-frequency function tokens or tokens that frequently co-occur with the corresponding relations. Although most transformer-based pre-trained language models (PLMs) are expected to maximize the probability of current word Y given its context X, which is represented by conditional distribution $P(Y|X)$, they have instead learned $P(Y|X, A)$, where A denotes the access to the sampling process. The selection bias results in spurious correlations between high-frequency function tokens and current tokens. Specifically, we explain the causal relationships between variables during pre-training PLMs and represent it in a causal directed acyclic graph (DAG) as shown in Figure 5.

As the high-frequency function words H possess grammatical meaning (e.g.,'.' and 'the'), they are more possible to be sampled either in training corpus or context, while other words U are relatively less likely to access the sampling process or context. The phenomenon is represented by $H \rightarrow A$ and $U \rightarrow A$, where directed edges denote causal relationships between variables. However, the semantic meaning (word embedding) of the current word Y largely depends on the words carrying an explicit semantic meaning, which is depicted by $U \rightarrow Y$. In linguistics, content words contribute to the meaning of sentences where they occur, and function words express grammatical relationships among other words. Their combinations, which are implicit and hard to be intervened, form natural language expressions. The process can be described by $A \rightarrow X$, where A determines the word distribution over contexts.

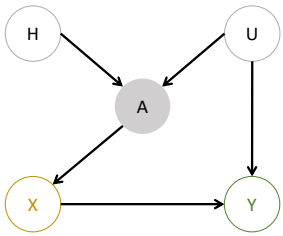

Figure 5: Causal graph of the sampling process.

Existing PLMs are pre-trained on a given corpus, indicating that A is given. Conditioning on A, the unconditionally independent variables H and U become dependent, which is described as collider bias (Pearl, 2009). Due to the causal relationship between U and Y, H and Y are spuriously correlated. That is to say, models tend to spuriously correlate high-frequency function words with any word, including word-level evidence which causes relations. Therefore, spurious correlations between high-frequency function words and relations are learned by models and represented in Figure 4.

Meanwhile, we can also observe spurious correlations between entity names and relations. Our analysis of the underlying causes is roughly the same as we mentioned before. We regard H as high-frequency entities from the relation-specific documents in Wikipedia, U as evidence words that causally determine the relation, Y as predictions, and X as documents. Given X and A, models learn spurious correlations between H and Y.

5.2 Attacks on the SOTA DocRE Models

In this section, we propose several RE-specific attacks to reveal the following facts: (1) The decision

Model	Mask.		ASA	SSA		EM	ES	ER	Val	HWE	Scratch
	P2N	UP	UP	P2N	UP	F1	F1	F1	F1	F1	F1
$\text{ATLOP}_{\text{BERT}}$ (Zhou et al., 2020)	**20.21**	**79.43**	**90.38**	6.47	93.46	6.39	6.08	14.16	61.09	57.69	40.56
$\text{ATLOP}_{\text{RoBERTa}}$ (Zhou et al., 2020)	16.51	82.98	90.42	3.85	96.02	27.29	7.35	17.50	63.18	58.43	42.12
$\text{DocuNet}_{\text{RoBERTa}}$ (Zhang et al., 2021)	16.49	83.19	91.48	2.82	97.17	8.62	8.08	18.55	63.91	59.58	42.78
$\text{SSAN}_{\text{RoBERTa}}$ (Xu et al., 2021)	13.68	85.48	91.23	1.73	98.26	35.41	6.09	22.72	62.08	58.37	48.74
$\text{EIDER}_{\text{RoBERTa}}$ (Xie et al., 2022)	14.24	85.36	92.78	2.12	97.88	**35.45**	**8.46**	**23.00**	64.28	60.62	**49.95**
$\text{KD}^{\dagger}_{\text{RoBERTa}}$ (Tan et al., 2022a)	10.77	88.69	95.46	**1.28**	**98.72**	29.74	7.57	20.35	**67.12**	**62.87**	45.82

Table 1: Results of different attacks. Model denoted by † is trained by extensive distantly supervised data. Val, HWE, and Scratch denote the validation set of DocRED, $\text{DocRED}_{\text{HWE}}$, and $\text{DocRED}_{\text{Scratch}}$. To observe the ratios of changed predictions in various attacks based on human-annotated word-level evidence, we propose P2N and UP. They denote the ratio of "negative predictions changed from positive predictions" to "original positive predictions", and "unchanged positive predictions" to "original positive predictions", respectively.

rules of models are largely different from that of humans. (2) Such a difference will severely damage the robustness and generalization ability of models: If a certain model always neglects the rationales in DocRE, it can hardly be aware of the tiny but crucial modifications on rationales. We introduce more details of our proposed attacks as follows.

Word-level Evidence Attacks. We present three kinds of attacks according to our proposed word-level evidence annotation: (1) *Masked word-level evidence attack* where all of the human-annotated word-level evidence (HWE) is directly masked; (2) *Antonym substitution attack (ASA)* where a word in HWE is replaced by its antonyms; (3) *Synonym substitution attack (SSA)* where a word in HWE is replaced by its synonyms. Since some evidence words do not have antonyms or synonyms in WordNet (Miller, 1995), we attack the rest of the words in HWE. Note that we only attack the HWE of those relation facts that have a single reasoning path to make sure our antonym/synonym substitution will definitely change/keep the original label. Specifically, in ASA, we first select the first suitable word in HWE that either possesses its antonym in WordNet or belongs to different forms of the verb "be". We generate the opposite meaning either by adding "not" after the "be" verbs or substituting the word with its antonym. In SSA, the first suitable word in HWE will be replaced by its synonyms. We conduct ASA and SSA on 2,002 and 5,321 relational facts, respectively.

The results of the three kinds of attacks are shown in Table 1. Under the masked word-level evidence attacks, the evidence supporting the relational facts is removed. The relations between entity pairs are supposed not to exist. However, we can observe that, as to the best performance, no more than 21% of predictions is even changed. Models still predict the same relations even if they are erased, which leads to at least a 79% decline in the performance of models. As to ASA, the semantic meanings of evidence are changed to the opposite. Models are expected to alter their predictions. However, the SOTA models alter no more than 10% predictions after the attack, which indicates that the performance of models will sharply drop by at least 90% under ASA. The results of SSA are roughly the same as ASA. According to the experimental results of previous attacks, we can attribute the good performance of models under SSA to the fact that models are hardly aware of rationales. All three kinds of attacks confirm the conclusion that the decision rules of models are largely different from that of humans. The difference severely damages the robustness of models.

Entity Name Attacks As shown in Section 5.1, we observe that models rely largely on tokens in entities. To further investigate the extent to which models depend on entity names to improve their performance, we design a few attacks to exhibit their bottleneck. We propose (1) mask entity attack (EM) where we directly mask all entity names, (2) randomly shuffled entity attack (ER) where we randomly permute the names of entities in each document, and (3) out-of-distribution (OOD) entity substitution attack (ES) where we use entity names that have never occurred in training data to substitute the entity names in an input document. As shown in Table 1, we observe significant declines in the F1 scores from all models. The experimental results are shown in Table 1. The most significant performance decline occurs when attacking $\text{KD}_{\text{RoBERTa}}$ by ES, where the F1-score drops from 67.12% to 7.57%.

The results of entity name attacks show that models spuriously correlate entity names with the final

predictions. In other words, they make predictions according to entity names. The poorer the performance, the more spurious correlations are learned. The differences are: (1) EM removes original entity name information to detect spurious correlations; (2) ER modifies original entity name information to attack the learned spurious correlations, making them misleading to further test the robustness of models; (3) OOD-ES removes original entity name information and introduces new OOD entity name information, evaluating the generalization ability of models on tackling the unseen entity name information without the help of spurious correlations.

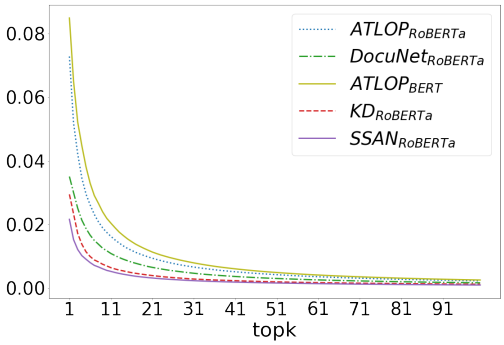

Figure 6: MAP curves of different models.

5.3 Evaluation Metric

In Section 5.2, we demonstrate that the decision rules of models should approach that of humans to improve the understanding and reasoning capabilities of models. The desiderata of the capabilities and the similar conclusions are also presented in other NLP tasks (Jia and Liang, 2017; Wang et al., 2022). However, how do we measure the extent to which models possess these capabilities? In other words, how to measure the distance between the decision rules of models and that of humans? In previous work, they calculate F1-score over the evidence sentences. Models are trained to recognize the corresponding right evidence sentences when they extract a relational fact. Despite the plausible process, the recognized holistic evidence sentences fail to provide fine-grained word-level evidence, resulting in unfaithful observations discussed in Section 3.1. Furthermore, models' performance of predicting evidence sentences can not represent their understanding and reasoning capabilities: the blackbox process of learning how to predict evidence may introduce other new problems in the newly learned decision rules.

To solve the issue, we introduce mean average precision (MAP) (Zhu, 2004) to evaluate the performance of models and explain their reliability. We also visualize the MAP values of the models.

MAP is a widely adopted metric to evaluate the performance of models, including Faster R-CNN (Ren et al., 2015), YOLO (Redmon et al., 2016), and recommender systems (Ma et al., 2016). We note that evaluating recommender systems and measuring the capabilities of models share a common background. Intuitively, we can consider "the human-annotated evidence words" as "the relevant items for a user", and "the most crucial words considered by a certain model" as "the recommended items of a recommender system". Consequently, given top K words with the highest attribution values, the formula of MAP over T relational facts can be written by,

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

where $\mathbf{1}_t(i)$ denotes the indicator function of the i-th important word for predicting the t-th relational fact. The output value of $\mathbf{1}_t(i)$ equals 1 if the word is in the human-annotated word-level evidence. Else it equals 0. The selection of K, similar to the evaluation metrics in recommender systems, depends on the demand of RE practitioners and is often set to 1, 10, 50, and 100. Also, we can select all the possible values of K to form a MAP curve and measure the AUC to holistically evaluate the understanding ability of models. For each relational fact, words "recommended" by models will be evaluated according to 1) how precise they perform the human-annotated word-level evidence, and 2) the "recommending" order of these important words determined by their attribution values. Based on MAP, we measure the extent to which the decision rules of models differ from that of humans. Due to the mechanism of EIDER where documents and the predicted sentences from documents are combined together to predict by truncation, it is impractical to attribute EIDER by gradient-based methods. We compute MAP for other SOTA models. The results are shown in Figure 6. We can observe that the MAP values of SOTA models are all below 10%, which is far less than the average level of normal recommender systems. Obviously, existing models fail to understand the documents as humans do, which explains the reason why they

are vulnerable to our proposed attacks.

In this section, we use MAP to evaluate to which extent a model makes decisions like a human, which indicates the brittleness and robustness of a model. Models can explore many ways to achieve a good performance on the test set (represented by F1 score), including greedily absorbing all correlations found in data, recognizing some spurious patterns, etc., but MAP will tell us which model is trustworthy or robust and can be deployed in real-world applications.

5.4 Discussion

In this section, we discuss the connections between some experimental results to give some instructive advice. First, we can observe that for the models whose MAP value is larger, their performance under word-level evidence-based attacks will be better. MAP curve reflects the extent to which models possess understanding and reasoning abilities. As shown in Figure 6, the various extents can be described from high to low by $\text{ATLOP}_{\text{BERT}} > \text{ATLOP}_{\text{RoBERTa}} > \text{DocuNet}_{\text{RoBERTa}} > \text{KD}_{\text{RoBERTa}} \approx \text{SSAN}_{\text{RoBERTa}}$, which is consistent with the performance levels under mask word-level evidence attack and antonym substitution attack represented from high to low by $\text{ATLOP}_{\text{BERT}} > \text{ATLOP}_{\text{RoBERTa}} > \text{DocuNet}_{\text{RoBERTa}} > \text{KD}_{\text{RoBERTa}} \approx \text{SSAN}_{\text{RoBERTa}}$. Furthermore, if the decision rules of models largely differ from that of humans (MAP value is small), it will be ambiguous to identify which kind of attack the models will be vulnerable to. According to the results in Table 1, the performance of models are irregular under entity name attacks. The underlying causes can be any factors that can influence the training effect on a model.

Although training on extensive distantly supervised data can lead to the performance gain on the validation set of DocRED and $\text{DocRED}_{\text{HWE}}$, it also renders the poor understanding and reasoning capabilities of $\text{KD}_{\text{RoBERTa}}$ according to Figure 6, which makes it be the most vulnerable model under mask word-level evidence attack and antonym substitution attack. As shown in Table 1, the generalization ability of $\text{KD}_{\text{RoBERTa}}$ is also weakened when compared with $\text{EIDER}_{\text{RoBERTa}}$ on $\text{DocRED}_{\text{Scratch}}$, which does not use any extra training data and predicts through evidence sentences annotated by humans. $\text{EIDER}_{\text{RoBERTa}}$ simultaneously enhances the performance, generalization ability, and robustness of models. We can observe its stronger robustness under entity name attacks, outstanding performance on the validation set of DocRED and $\text{DocRED}_{\text{HWE}}$, and stronger generalization ability on $\text{DocRED}_{\text{Scratch}}$. The success of $\text{EIDER}_{\text{RoBERTa}}$ indicates that rationales considered by humans are of the essence in DocRE.

All the results indicate that guiding a model to learn to predict by the evidence of humans can be the essential way to improve the robustness of models, thus making models trustworthy in real-world applications.

6 Limitation

In this paper, we propose $\text{DocRED}_{\text{HWE}}$ and introduce a new metric to select the most robust and trustworthy model from those well-performed ones in DocRE. However, all data in DocRED are sampled from Wikipedia and Wikidata, which indicates that training and test data in DocRED can be identically and independently distributed (i.i.d. assumption). The i.i.d. assumption impedes our demonstration of the intuition: A model with a higher MAP will obtain a higher F1 score on the test set. Due to the i.i.d. assumption, models can succeed in obtaining a higher F1 score by greedily absorbing all correlations (including spurious correlations) in the training data. To strictly demonstrate the intuition, we need a test set that exhibits different and unknown testing distributions. In addition, expanding the research scope to a cleaner Re-DocRED and analyzing the role of unobservable wrong labels are also crucial and interesting ideas. We leave them as our future work.

7 Conclusion

Based on our analysis of the decision rules of existing models on DocRE and our annotated word-level evidence, we expose the bottleneck of the existing models by our introduced MAP and our proposed RE-specific attacks. We also extract some instructive suggestions by exploring the connections between the experimental results.

We appeal to future research to take understanding and reasoning capabilities into consideration when evaluating a model and then guide models to learn evidence from humans. Based on proper evaluation and guidance, significant development can be brought to the document-level RE, where the performance, generalization ability, and robustness of models are more likely to be improved.

Acknowledgement

This work is partially supported by funds from Arcplus Group PLC (Shanghai Stock Exchange: 600629).

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. 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.

Jinlan Fu, Pengfei Liu, and Qi Zhang. 2020. Rethinking Generalization of Neural Models: A Named Entity Recognition Case Study. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7732–7739.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

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, Dublin, Ireland. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

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*, 2016.

Frederick Liu and Besim Avci. 2019. Incorporating Priors with Feature Attribution on Text Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, Florence, Italy. Association for Computational Linguistics.

Jian Liu, Yufeng Chen, and Jinan Xu. 2022. Saliency as Evidence: Event Detection with Trigger Saliency Attribution. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 4573–4585, Dublin, Ireland. Association for Computational Linguistics.

Hao Ma, Xueqing Liu, and Zhihong Shen. 2016. User Fatigue in Online News Recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1363–1372, Montréal Québec Canada. International World Wide Web Conferences Steering Committee.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the Model Understand the Question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with Latent Structure Refinement for Document-Level Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557.

Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-Sentence *N* -ary Relation Extraction with Graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Yu-Ming Shang, Heyan Huang, and Xian-Ling Mao. 2022. OneRel:Joint Entity and Relation Extraction with One Module in One Step. *arXiv:2203.05412 [cs]*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

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, Dublin, Ireland. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States. Association for Computational Linguistics.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage Joint Extraction of Entities and Relations Through Token Pair Linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488.

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 *International Conference on Research in Computational Molecular Biology*, pages 272–284.

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

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, Dublin, Ireland. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity Structure Within and Throughout: Modeling Mention Dependencies for Document-Level Relation Extraction. In *AAAI*, pages 14149–14157.

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.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02*, volume 10, pages 71–78, Not Known. Association for Computational Linguistics.

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.

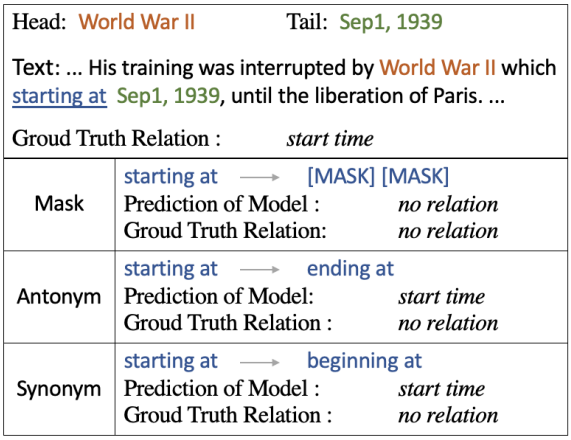

Figure 7: An example for the three kinds of attacks.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *IJCAI*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2020. Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling. In *AAAI*, pages 14612–14620.

Mu Zhu. 2004. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2(30):6.

A Details of Attacks

We give an example to illustrate our proposed three kinds of word-level evidence attacks. The example is shown in Figure 7

B Annotation Errors in DocRED

We provide the details of all our corrected errors in our selected 699 documents from the validation set of the original DocRED. All the error descriptions are shown in Table 2, Table 3, and Table 4. Annotators correct three kinds of annotation errors, which are exhibited in Table 5 and Table 6. "Err.1" denotes relation type error where annotators wrongly annotate a relation type between an entity pair. "Err.2" denotes insufficient evidence error where an annotated relation can not be inferred from the corresponding document. "Err.3" denotes evidence error where the sentence-level evidence of a relation is wrongly annotated.

Document Title	Rel.	Error Description
The Time of the Doves	2	The relation can only be inferred by the first sentence.
The Time of the Doves	4	The relation is not P150.
Hélé Béji	8	No evidence can be found for this relation.
Hélé Béji	1	We can't infer relation P569 from the first evidence sentence.
Ne crois pas	1	The only evidence sentence of P27 is the sentence 5 instead of 0.
Ne crois pas	9	The only evidence sentence of P27 is the sentence 5 instead of 2.
Ne crois pas	14	The only evidence sentence of P1344 is the sentence 7 instead of 2.
Asian Games	5	No evidence can be found for this relation.
Asian Games	7	No evidence can be found for this relation.
The Longest Daycare	10	The second sentence does not clearly indicate that David is the director and only the third sentence indicates it, so the evidence is [0,3]
The Longest Daycare	28	the zeroth sentence can't infer that Simpsons are from the United States and Only the seventh sentence indicates it, so the evidence is [0,7]
South Gondar Zone	1	P150 can not be inferred, evidence is null, can't find evidence
South Gondar Zone	4	P17 can not be inferred according to the given document.
South Gondar Zone	16	P403 can not be inferred according to the given document.
South Gondar Zone	3	Evidence of the third relation(P27) in labels is [0,1] instead of [0,1,2]
Milton Friedman ...	1	"Evidence of the first relation(P31) in labels is [0] instead of [0,3]"
Milton Friedman ...	8	Evidence of the eighth relation(P108) in labels is [8] instead of [7,8]
Fedor Ozep	2	Evidence of the second relation(P20) in labels is [6] instead of [0,6]
TY.O	1	Evidence of P264 is [1] instead of [0,1]
TY.O	3	Evidence of P175 is [3] instead of [0, 3, 4]
TY.O	10	Evidence of P175 is [0] instead of [0, 4]
TY.O	13	Evidence of 162 is [0,3] instead of [0, 3, 4]
TY.O	14	Evidence of P175 is [0,3] instead of [0, 3, 4]
TY.O	20	Evidence of P175 is [0,3] instead of [0, 3, 4]
TY.O	29	Evidence of P175 is [0,3] instead of [0, 3, 4]
Front of Islamic ...	1	Evidence of P1412 is [0,2] instead of [0, 2, 4]
Front of Islamic ...	2	Evidence of P1412 is [0,2] instead of [0, 2, 4]
Front of Islamic ...	3	Evidence of P37 is [0,3] instead of [0, 3, 4]
Front of Islamic ...	4	Evidence of P1412 is [0,2] instead of [0, 2, 4]
Front of Islamic ...	5	Evidence of P1412 is [0,3] instead of [0, 3, 4]
Rufus Carter	7	P131 represents "located in the administrative territorial entity",but it can not be inferred according to the given document.
Rufus Carter	8	P150 can not be inferred according to the given document.
Smoke Break	1	Evidence of P577 is [1] instead of [1,8]
Smoke Break	2	Evidence of P264 is [1] instead of [1,2]
Smoke Break	3	Evidence of P162 is [2] instead of [0,2]
Bambi II	6	P17 can not be inferred according to the given document.
Bambi II	8	P272 can not be inferred according to the given document.
Bambi II	13	P272 can not be inferred according to the given document.
Bambi II	15	P272 can not be inferred according to the given document.
Assassin's Creed Unity	1	P178 can not be inferred according to the given document.
Assassin's Creed Unity	1	P178 can not be inferred according to the given document.
Assassin's Creed Unity	16	P577 can not be inferred according to the given document.
Assassin's Creed Unity	13	P179 can not be inferred according to the given document.
Assassin's Creed Unity	3	P123 can not be inferred according to the given document.
Mehmet Çetingöz	1	P17 can not be inferred according to the given document.
Mehmet Çetingöz	2	P17 can not be inferred according to the given document.
Mehmet Çetingöz	9	P17 can not be inferred according to the given document.
Mehmet Çetingöz	12	P17 can not be inferred according to the given document.
Mehmet Çetingöz	10	P17 can not be inferred according to the given document.
Baltimore and ...		Evidence of P17 is [0,2,3] instead of [0,2]
Baltimore and ...		Evidence of P279 is [2] instead of [2,4]
Dante Alighieri Society	6	Evidence of P571 is [0,2] instead of [0,2,5]
Osaka Bay	25	Evidence of P17 is [0,3,11] instead of [0,11]
Osaka Bay	36	Evidence of P17 is [0,3,11] instead of [0,11]
Osaka Bay	38	Evidence of P17 is [0,3,11] instead of [0,11]
Liang Congjie	8	relation can not be inferred from the context

Table 2: Wrong annotations in the original DocRED.

Document Title	Rel.	Error Description
University (album)	18	Evidence of P264 should be [3,4]
University (album)	19	Evidence of P175 should be [3]
University (album)	24	Evidence of P527 should be [3]
University (album)	25	Evidence of P475 should be [0]
Lappeenranta	2	Evidence of P131 is [1,2,3] instead of [1,3]
Lappeenranta	12	Evidence of P17 is [0,2,4,5,7,9] instead of [0,2,4,7,9]
Lappeenranta	13	Evidence of P131 is [1] instead of [0,1]
Lappeenranta	18	Evidence of P131 is [1,3] instead of [1,2,3]
Ali Abdullah Ahmed	4	Evidence of P3373 is [6] instead of [3,6]
Ali Abdullah Ahmed	8	Evidence of P3373 is [6] instead of [3,6]
Ali Abdullah Ahmed	9	Evidence of P570 is [7] instead of [6,7]
Joseph R. Anderson	9	P571 can not be inferred according to the given document.
Ramblin' on My Mind	1	Evidence of P175 is [5] instead of [0,2]
Ramblin' on My Mind	2	P86 can not be inferred according to the given document.
Christopher Franke	3	Evidence of P463 is [1,3,4,5] instead of [0,1,3,5]
Christopher Franke	4	P159 can not be inferred according to the given document.
Christopher Franke	5	P577 can not be inferred according to the given document.
Statue of Jan Smuts	3	Evidence of P27 is [5] instead of [4,5]
Statue of Jan Smuts	4	Evidence of P27 is [5] instead of [4,5]
Robert Taylor	1	Evidence of P108 is [1] instead of [0,1]
Robert Taylor	2	Evidence of P27 is [2] instead of [4,5]
Robert Taylor	3	Evidence of P27 is [3] instead of [4,5]
Robert Taylor	4	Evidence of P27 is [4] instead of [4,5]
Sycamore Canyon	1	P17 can not be inferred according to the given document.
Amos Hochstein	9	P194 can not be inferred according to the given document.
Paul Pfeifer	3	P69 can not be inferred according to the given document.
Mega Man Zero	8	P155 can not be inferred, Virtual Console is Wii U
Soldier (song)	1	Evidence of P577 is [1] instead of [0,1]
Soldier (song)	3	Evidence of P495 is [2] instead of [0,2]
Gloria Estefan Albums Discography	4	P156 can not be inferred. Let It Loose and Cuts Both Ways are two albums published one after another instead of two songs from an album. They are independent of each other. There is no evidence in the context.
Anthony G. Brown	3	Evidence of P27 is [0,4] instead of [0,3].
Harbour Esplanade	3	P17 can not be inferred according to the given document.
Harbour Esplanade	5	P17 can not be inferred according to the given document.
Harbour Esplanade	6	P17 can not be inferred according to the given document.
Henri de Buade	3	The relation between France and New France is colony instead of P495.
The Reverent Wooing of Archibald	5	P577 should be P580.
This Little Girl of Mine	6	The third sentence should be removed from the evidence of P136.
This Little Girl of Mine	9	The zeroth sentence should be removed from the evidence of P175.
This Little Girl of Mine	13	The zeroth sentence should be removed from the evidence of P175.
This Little Girl of Mine	15	The zeroth sentence should be removed from the evidence of P264, it only refers to the name of the head entity.
This Little Girl of Mine	19	The zeroth sentence should be removed from the evidence of P175 it only refers to the name of the tail entity.
This Little Girl of Mine	20	"The zeroth sentence should be removed from the evidence of P175, it only refers to the name of the performer and can't infer the relation between two sides."
Ali Akbar Moradi	1	Evidence of P569 should be [0].
Ali Akbar Moradi	2	The zeroth sentence should be removed from the evidence of P19, it only refers to the name of the head entity.
Ali Akbar Moradi	3	The zeroth sentence should be removed from the evidence of P27, it only refers to the name of the head entity.
I Knew You Were Trouble	2	The zeroth sentence should be removed from the evidence of P264, because no words related to two entities can be found in it.
I Knew You Were Trouble	4	The zeroth sentence should be removed from the evidence of P175 it only refers to the name of the head entity.
I Knew You Were Trouble	5	The zeroth sentence should be removed from the evidence of P577 it only refers to the name of the head entity.

Table 3: Wrong annotations in the original DocRED.

Document Title	Rel.	Error Description
I Knew You Were Trouble	6	The zeroth sentence should be removed from the evidence of P495 it only refers to the name of the head entity.
I Knew You Were Trouble	7	The zeroth sentence should be removed from the evidence of P264 it only refers to the name of the head entity.
I Knew You Were Trouble	8	The zeroth sentence should be removed from the evidence of P162 it only refers to the name of the head entity.
I Knew You Were Trouble	9	The zeroth sentence should be removed from the evidence of P361 it only refers to the name of the head entity.
Mohammed Abdel Wahab	6	P86 can not be inferred according to the given document.
Mohammed Abdel Wahab	8	P86 can not be inferred according to the given document.
Mohammed Abdel Wahab	10	P86 can not be inferred according to the given document.
Elbląg County	5	Evidence of P150 is [0,2] instead of [0,2,3].
The Crazy World of Arthur Brown (album)	1	P264 represents “brand and trademark associated with the marketing of subject music recordings and music videos”, but here the head entity is the same name as music, instead of a music album.
The Crazy World of Arthur Brown (album)	6	P264 represents “brand and trademark associated with the marketing of subject music recordings and music videos”, but here the head entity is the same name as music, instead of a music album."
The Crazy World of Arthur Brown (album)	7	P264 represents “brand and trademark associated with the marketing of subject music recordings and music videos”, but here the head entity is the same name as music, instead of a music album.
The Crazy World of Arthur Brown (album)	8	P264 represents “brand and trademark associated with the marketing of subject music recordings and music videos”, but here the head entity is the same name as music, instead of a music album.
The Crazy World of Arthur Brown (album)	9	P264 represents “brand and trademark associated with the marketing of subject music recordings and music videos”, but here the head entity is the same name as music, instead of a music album.
Flag of Prussia	1	Evidence of P155 is [0] instead of [2,4].
Flag of Prussia	3	P155 should be P6.
Flag of Prussia	7	Evidence of P156 is [0] instead of [2,4].
Flag of Prussia	11	P156 represents “immediately following item in a series of which the subject is a part”, but here both entities are the same.
John Alexander Boyd	11	Evidence of P17 is [0] instead of [0,5].
John Alexander Boyd	12	Evidence of P17 is [0] instead of [0,5,6].
Municipal elections in Canada	5	Evidence of P17 is [8] instead of [8,11].
Municipal elections in Canada	7	Evidence of P131 is [11] instead of [0,8,11].
House of Angels	7	Evidence of P495 is [0,8] instead of [0,6,8].
William James Wallace	7	Evidence of P17 is [0,1] instead of [1,3].
William James Wallace	8	P17 can not be inferred according to the given document.
William James Wallace	10	P27 can not be inferred according to the given document.
William James Wallace	11	P17 can not be inferred according to the given document.
Black Mirror (song)	7	Evidence of P264 is [2] instead of [0,2].
Michael Claassens	5	Evidence of P264 is [4] instead of [0,4].
Michael Claassens	12	Evidence of P264 is [6] instead of [0,6].
Lark Force	13	the zeroth sentence can’t infer that HMAT Zealandia is from Australia.
Washington Place (West Virginia)	9	the zeroth sentence can’t infer that Annie Washington is from the United States.
Battle of Chiari	2	Evidence of P276 is [0,2] instead of [0,3].
Battle of Chiari	6	Evidence of P607 is [1,2] instead of [1].
Woodlawn, Baltimore County, Maryland	18	Evidence of P131 is [0,5] instead of [0,4,5].
Wagner–Rogers Bill	1	Evidence of P27 is [0,1] instead of [0].

Table 4: Wrong annotations in the original DocRED.

Document Title	Rel.	Err. 1	Err. 2	Err. 3
The Time of the Doves	2			✓
The Time of the Doves	4	✓		
Hélé Béji	8		✓	
Hélé Béji	1			✓
Ne crois pas	1			✓
Ne crois pas	9			✓
Ne crois pas	14			✓
Asian Games	5			✓
Asian Games	7			✓
The Longest Daycare	10			✓
The Longest Daycare	28			✓
South Gondar Zone	1		✓	✓
South Gondar Zone	4		✓	
South Gondar Zone	16		✓	
South Gondar Zone	3			✓
Milton Friedman ...	1			✓
Milton Friedman ...	8			✓
Fedor Ozep	2			✓
TY.O	1			✓
TY.O	3			✓
TY.O	10			✓
TY.O	13			✓
TY.O	14			✓
TY.O	20			✓
TY.O	29			✓
Front of Islamic ...	1			✓
Front of Islamic ...	2			✓
Front of Islamic ...	3			✓
Front of Islamic ...	4			✓
Front of Islamic ...	5			✓
Rufus Carter	7		✓	
Rufus Carter	8		✓	
Smoke Break	1			✓
Smoke Break	2			✓
Smoke Break	3			✓
Bambi II	6		✓	
Bambi II	8		✓	
Bambi II	13		✓	
Bambi II	15		✓	
Assassin's Creed Unity	1		✓	
Assassin's Creed Unity	16		✓	
Assassin's Creed Unity	13		✓	
Assassin's Creed Unity	3		✓	
Mehmet Çetingöz	1		✓	
Mehmet Çetingöz	2		✓	
Mehmet Çetingöz	9		✓	
Mehmet Çetingöz	12		✓	
Mehmet Çetingöz	10		✓	
Baltimore and ...				✓
Baltimore and ...				✓
Dante Alighieri Society	6			✓
Osaka Bay	25			✓
Osaka Bay	36			✓
Osaka Bay	38			✓
Liang Congjie	8		✓	
University (album)	18			✓
University (album)	19			✓
University (album)	24			✓
University (album)	25			✓
Lappeenranta	2			✓
Lappeenranta	12			✓
Lappeenranta	13			✓
Lappeenranta	18			✓
Ali Abdullah Ahmed	4			✓
Ali Abdullah Ahmed	8			✓
Ali Abdullah Ahmed	9			✓
Joseph R. Anderson	9		✓	
Ramblin' on My Mind	1			✓
Ramblin' on My Mind	2		✓	

Table 5: The category of each error in the original DocRED.

Document Title	Rel.	Err. 1	Err. 2	Err. 3
Christopher Franke	3			✓
Christopher Franke	4		✓	
Christopher Franke	5		✓	
Statue of Jan ...	3			✓
Statue of Jan ...	4			✓
Robert Taylor	1			✓
Robert Taylor	2			✓
Robert Taylor	3			✓
Robert Taylor	4			✓
Sycamore Canyon	1		✓	
Amos Hochstein	9		✓	
Paul Pfeifer	3		✓	
Mega Man Zero	8		✓	
Soldier (song)	1			✓
Soldier (song)	3			✓
Gloria Estefan ...	4		✓	
Anthony G. Brown	3			✓
Harbour Esplanade	3		✓	
Harbour Esplanade	5		✓	
Harbour Esplanade	6		✓	
Henri de Buade	3	✓		
The Reverent ...	5	✓		
This Little ...	6			✓
This Little ...	9			✓
This Little ...	13			✓
This Little ...	15			✓
This Little ...	19			✓
This Little ...	20			✓
Ali Akbar Moradi	1			✓
Ali Akbar Moradi	2			✓
Ali Akbar Moradi	3			✓
I Knew You ...	2			✓
I Knew You ...	4			✓
I Knew You ...	5			✓
I Knew You ...	6			✓
I Knew You ...	7			✓
I Knew You ...	8			✓
I Knew You ...	9			✓
Mohammed A. W.	6		✓	
Mohammed A. W.	8		✓	
Mohammed A. W.	10		✓	
Elbląg County	5			✓
The Crazy World ...	1		✓	
The Crazy World ...	6		✓	
The Crazy World ...	7		✓	
The Crazy World ...	8		✓	
The Crazy World ...	9		✓	
Flag of Prussia	1			✓
Flag of Prussia	3	✓		
Flag of Prussia	7			✓
Flag of Prussia	11		✓	
John Alexander Boyd	11			✓
John Alexander Boyd	12			✓
Municipal elections ...	5			✓
Municipal elections ...	7			✓
House of Angels	7			✓
William James Wallace	7			✓
William James Wallace	8		✓	
William James Wallace	10		✓	
William James Wallace	11		✓	
Black Mirror (song)	7			✓
Michael Claassens	5			✓
Michael Claassens	12			✓
Lark Force	13			✓
Washington Place	9			✓
Battle of Chiari	2			✓
Battle of Chiari	6			✓
Woodlawn, Baltimore ...	18			✓
Wagner–Rogers Bill	1			✓

Table 6: The category of each error in the original DocRED.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☒ A2. Did you discuss any potential risks of your work?
Our work can cause no potential risks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ 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.
No response.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5

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?
5

☑ 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?
5

☑ 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.)?
55

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☑ 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.?
3

☑ 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)?
3

☑ 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?
3

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
3

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3