Ontology-Driven Structural Regularization for Document-Level Relation Extraction

Laura Menotti[1], **Stefano Marchesin**[1], and **Gianmaria Silvello**[1]
[1]Department of Information Engineering, University of Padua, Italy
{name.surname}@unipd.it

Abstract

Document-Level Relation Extraction (DocRE) relies heavily on costly manually annotated datasets, while large distant supervision resources such as *DocRED distant* remain under-exploited due to noise. We show that a critical yet overlooked source of noise lies in structural inconsistencies within relational triples, including violations of ontology constraints and logical contradictions.

We introduce an ontology-driven framework to quantify and enforce structural consistency in DocRE datasets. Our analysis reveals substantial structural noise in DocRED distant and demonstrates that such inconsistencies propagate to model predictions. Enforcing structural well-formedness during training significantly reduces logical contradictions and consistently improves generalization performance. These findings establish structural consistency as a missing axis of supervision in DocRE and highlight structural regularization as an effective strategy for leveraging distant data at scale.

1 Introduction

Document-Level Relation Extraction (DocRE) aims to identify all semantic relations holding between entity pairs within a full document. By extracting relational triples, DocRE enables the construction of Knowledge Graphs (KGs), which provide structured, machine-readable knowledge to support downstream tasks such as information retrieval, question answering, data mining, and recommendation systems (Dong, 2023). Building accurate KGs therefore depends on reliable DocRE models trained on high-quality data.

The reference benchmark for general-domain DocRE is DocRED (Yao et al., 2019), constructed from Wikipedia articles. DocRED includes both manually annotated datasets and a large distantly supervised dataset. The manually curated split follows a recommend-revise paradigm and contains 5,053 documents, divided into 3,053 for training and 1,000 each for development and testing. In contrast, the DocRED distant dataset comprises 101,873 documents, annotated via Distant Supervision (DS) by aligning Wikidata triples with Wikipedia text and leveraging BERT-based named entity annotations.

While distant supervision enables large-scale data construction (Mintz et al., 2009), it relies on the strong assumption that if a KG asserts a relation between two entities, then every document mentioning the pair expresses that relation. This assumption adds substantial noise. In DocRE, the issue is worse because entities co-occurring in a document are less likely to express a true relation than those in the same sentence. As a result, most state-of-the-art models rely mainly on manually annotated data or use distant data only for pre-training or distillation (Tan et al., 2022a).

Recent analyses have further revealed annotation issues even in the manually curated splits. In particular, around 65% of positive relations were missing in the original DocRED annotations (Huang et al., 2022), leading to the release of ReDocRED, a revised version of the training and development datasets (Tan et al., 2022b). Although existing denoising approaches attempt to mitigate DS noise through uncertainty estimation (Sun et al., 2023) or active learning for long-tail relations (Menotti et al., 2026), they focus primarily on semantic correctness and model-driven pseudo-label refinement.

In contrast to model-centric denoising strategies, we adopt a *data-centric* approach, focusing on improving the structural quality of training data itself. We argue that beyond semantic noise, DocRE datasets suffer from *structural inconsistencies* in the extracted triples, such as violations of ontology constraints, missing inverse relations, asymmetric contradictions, and cardinality conflicts.[1] These

[1]We use the term structural consistency to denote the well-formedness of RDF triples under ontology-imposed constraints, rather than full ontology satisfiability.

inconsistencies represent a previously overlooked source of supervision noise that might propagate during model training.

Concretely, we operationalize structural consistency by leveraging the formal semantics of the Web Ontology Language (OWL) to define explicit structural constraints over entity types and relations. Our methodology evaluates the *well-formedness* of relational triples under ontology axioms and quantifies structural violations in both DocRED and Re-DocRED, with particular emphasis on the DocRED distant dataset.

We show that structural noise is substantially more prevalent in distantly supervised data than in manually curated datasets. In particular, compared to the ReDocRED manual training split, the DocRED distant dataset contains three times more invalid triples, nearly five times more missing inverse relations, and twice as many asymmetric inconsistencies. These findings suggest that structural inconsistencies constitute a systematic source of noise that can propagate during training, impairing a model's ability to learn and generate well-formed triples. To mitigate this effect, we introduce a lightweight, model-agnostic *pre-processing pipeline* that enforces structural consistency prior to training. The pipeline cleans and augments the dataset without altering model architectures, loss functions, or inference procedures, ensuring seamless integration with existing DocRE systems – whether sequence-based or graph-based.

We show that enforcing ontological consistency acts as a form of structural regularization, improving both prediction coherence and generalization. Our main contributions are: (1) We propose an ontology-driven framework that uses OWL constraints to rigorously diagnose ontology-level structural consistency in DocRE datasets; (2) We deliver the first in-depth quantitative study of structural inconsistencies in DocRED distant and its denoised variants, uncovering extensive, previously unreported violations; (3) We show that enforcing structural well-formedness as a pre-processing step markedly reduces logical contradictions in model predictions and consistently boosts generalization performance across state-of-the-art DocRE models (Zhou et al., 2021; Ma et al., 2023).

We release the complete analysis and cleaning framework as open-source.[2] The toolkit is compatible with any dataset formatted in the DocRED schema and includes the ontology-based rules we developed, while allowing for the definition of other custom constraints.

The rest of this paper is organized as follows. Section 2 reviews prior work on distant denoising and constrained Relation Extraction (RE). Section 3 defines the structural rules used in our analysis. Section 4 assesses the structural consistency of DocRED and ReDocRED and examines the main causes of invalid triples. Section 5 studies how structural noise affects DocRE models during training. Finally, Section 6 concludes the paper.

2 Related Work

Distant Supervision (DS) enables large-scale relation extraction by aligning entity pairs in text with triples from a reference KG (Mintz et al., 2009). While efficient, this paradigm introduces noisy supervision due to the heuristic alignment between text and KG facts.

To address this issue, several works have proposed denoising strategies for DocRE. Xiao et al. (2020) introduced a multi-task framework incorporating mention-entity matching and fact alignment. Sun et al. (2023) proposed uncertainty-guided label denoising using Monte Carlo dropout to filter low-confidence pseudo-labels. Menotti et al. (2026) developed an active learning pipeline tailored to improve long-tail relation predictions.

These approaches focus primarily on improving the semantic reliability of distant labels through model-driven refinement. In contrast, our work targets a complementary and largely unexplored dimension: structural consistency of relational triples under ontology-imposed constraints.

Logical rules have been explored in DocRE to enhance relation prediction. Ru et al. (2021) introduced a probabilistic framework that learns logic rules as latent variables and optimizes them via Expectation Maximization (EM). Zhang et al. (2024) proposed a Secondary Reasoning Framework that leverages first-stage predictions as contextual signals to infer additional relations among neighboring entities. Zhang et al. (2025) developed GREP, which models interdependencies between entity pairs to capture fine-grained reasoning patterns. These approaches incorporate logical or reasoning mechanisms within the prediction process, focusing on improving semantic inference. In contrast, our work addresses a complementary di-

[2] https://anonymous.4open.science/r/DocRE-StructuralConsistency-D01F/

mension: the structural well-formedness of triples under ontology-imposed constraints, independently of the model architecture or inference strategy.

Constraint-based approaches have also been explored in sentence-level RE to improve performance under distant supervision. Liu et al. (2014) leveraged entity type information to introduce fine-grained type constraints for relation prediction. Ji et al. (2017) incorporated Freebase entity descriptions as background knowledge to guide attention mechanisms and improve extraction accuracy. Liang et al. (2023) proposed a *constraint graph* constructed from entity type constraints, which provides an additional supervision signal within a constraint-aware attention module. While these methods integrate external knowledge to guide sentence-level inference, they have not been extended to DocRE, nor do they explicitly enforce ontology-level structural consistency.

3 Methodology

Given a document d and a set of entities $E = \{e_i\}_{i=1}^{n}$, each entity $e_i \in E$ is associated with a set of mentions $M_i = \{m_{i,j}\}_{j=1}^{k_i}$, where k_i denotes the number of occurrences of e_i in d. The task of DocRE is to predict a subset of relations from $\mathcal{R} \cup \{\mathsf{na}\}$ between entity pairs (e_s, e_o) for $s, o \in \{1, \ldots, n\}$ and $s \neq o$. Here, $\mathcal{R}$ is a pre-defined set of semantic relations, na represents the absence of a relation, and e_s, e_o denote the subject and object entities. [3] Rather than treating these predictions as independent classification decisions, we interpret DocRE as the construction of a KG focusing on local well-formedness conditions induced by OWL axioms. This perspective allows us to detect structural inconsistencies that are invisible to standard training objectives.

3.1 Invalid Triples

Each entity mention $m_{i,j}$ of an entity e_i is classified into a specific type from the set of named entity types $\mathcal{T} = \{$PER, ORG, LOC, TIME, NUM, MISC$\}$. For instance, the mention "*France*" is classified as LOC $\in \mathcal{T}$. Since all mentions in a set, say M_i, identify the same entity e_i, we can assume consistent typing across the set; thus, we use the first mention as representative, i.e., $t(e_i) = t(m_{i,1})$.

In this context, the entity pair (e_s, e_o) and their predicted relation $r \in \mathcal{R}$ form a Resource Description Framework (RDF) triple $\langle s, p, o \rangle$, where $s = e_s$ is the subject, $p = r$ is the predicate, and $o = e_o$ is the object. Beyond simple triple extraction, we can leverage OWL to define formal semantics over the set of entities E and relations $\mathcal{R}$. Each relation $r \in \mathcal{R}$ can be viewed as an `owl:ObjectProperty` with specific domain and range constraints. Specifically, for a triple $\langle e_s, r, e_o \rangle$ to be semantically valid, the entity types must satisfy: $t(e_s) \in \text{dom}(r)$ and $t(e_o) \in \text{ran}(r)$, where $\text{dom}(r), \text{ran}(r) \subseteq \mathcal{T}$ are the permitted subject and object types for relation r.

Triples that violate these constraints are considered *invalid triples* within the defined ontology. By treating $\mathcal{R}$ as a set of restricted properties rather than arbitrary labels, we can filter out noise and improve the logical consistency of the extracted document graph.

Each relation in DocRED is mapped to a wikidata property. [4] Thus, we leverage the Wikidata property descriptions to identify the domain (metadata field "`subject type constraint`") and the range of each property ("`value type constraint`"), if present. We also use the ReDocRED training dataset (Tan et al., 2022b) to extract the most popular entity types of the subject and object entities of each relation and include them in the domain and range. To avoid overly restrictive criteria, the domain and range of all relations include entity type "MISC". For instance, relation P19 (*place of death*) has domain "PER|MISC" and range "LOC|MISC".

Definition 3.1 (Invalid Triple)**.** Let $e_s, e_o \in E$ be a pair of entities with associated types $t(e_s), t(e_o) \in \mathcal{T}$ and $r \in \mathcal{R}$ be a relation; r is defined as **invalid** for the pair (e_s, e_o) if the subject entity type or the object entity type is not included in the domain or range of the property: $t(e_s) \notin \text{dom}(r) \quad \vee \quad t(e_o) \notin \text{ran}(r)$.

A triple $\langle e_s, r, e_o \rangle$ where relation r is invalid for the pair (e_s, e_o) is called an **invalid triple**. For instance, $\langle$Emily (PER), P19, 1987 (TIME)$\rangle$ is an invalid triple because the object entity type (TIME) is not allowed for relation P19 (*place of death*), as TIME $\notin$ ran(P19). The list of domain and range of each relation is reported in Tables 8 and 9 of Appendix D.

[3] We adopt the DocRED convention of *subject* and *object* to denote the first and second arguments of an RDF triple $\langle$s,p,o$\rangle$, not syntactic roles.

[4] For instance, metadata for property P26 (*spouse*) is available at: `https://www.wikidata.org/wiki/Property:P26`.

3.2 Missing Inverse Relations

Inverse properties in OWL allow for the formal definition of bidirectional relationships between entity instances. Within each document-level extracted graph, we adopt a local closed-world assumption with respect to inverse relations: if a relation r holds between two entities in the document, its declared inverse should also be present in the extracted graph. We identify a subset of relations $\mathcal{R}_{inv} \subset \mathcal{R}$ that possess an inverse property. Formally, for a relation $r \in \mathcal{R}$, there may exist a relation $r' \in \mathcal{R}$ such that $r' \equiv$ `owl:inverseOf` r. For instance, the relation $P1376$ (*capital of*) is the inverse of $P36$ (*capital*). To identify these pairs, we leverage the "`inverse property`" metadata field from the Wikidata properties description.

Definition 3.2 (Missing Inverse Relation)**.** Let $r, r' \in \mathcal{R}$ be two relations such that r' is declared as the `owl:inverseOf` r in the ontology. By the definition of `owl:inverseOf`, the presence of a triple $\langle e_s, r, e_o \rangle$ in a document d entails the existence of the inverse triple $\langle e_o, r', e_s \rangle$. Hence, a pair of entities (e_s, e_o) is said to have a **missing inverse relation** if:

$$\langle e_s, r, e_o \rangle \in d \quad \wedge \quad \langle e_o, r', e_s \rangle \notin d$$

Note that r and r' are not required to be distinct; if $r = r'$, the relation is defined as an `owl:SymmetricProperty`, where the predicate must hold in both directions. The list of inverse relations is reported in Table 10 of Appendix D.

3.3 Asymmetric Relation Inconsistencies

A relation $r \in \mathcal{R}$ is considered asymmetric if the existence of a relationship in one direction precludes its existence in the opposite direction. Formally, leveraging the "`inverse property`" metadata from Wikidata, we identify that the DocRED dataset contains only two symmetric relations: $P26$ (*spouse*) and $P3373$ (*sibling*), both of which satisfy $r \equiv$ `owl:SymmetricProperty`. All other relations $r \in \mathcal{R} \setminus \{P26, P3373\}$ are treated as asymmetric.

Definition 3.3 (Asymmetric Relation Violation)**.** Let $r \in \mathcal{R}$ be an `owl:AsymmetricProperty`. A triple $\langle e_s, r, e_o \rangle$ is defined as an **asymmetric triple**. A document d contains an **asymmetric relation violation** if both the triple and its direct inverse (using the same relation r) are present: $\langle e_s, r, e_o \rangle \in d \quad \wedge \quad \langle e_o, r, e_s \rangle \in d$, where $s \neq o$. Such a state represents a logical contradiction within the OWL definition of asymmetry, as the property cannot be its own inverse unless it is symmetric.

3.4 Relations Cardinality Violations

Given an entity subject e_s, the maximum cardinality of a relation r is defined as the largest possible number of distinct entity objects e_o such that triples of the form $\langle e_s, r, e_o \rangle$ can be instantiated. While domain knowledge initially suggests that a subset of 11 relations should be strictly functional ($k_r = 1$), empirical analysis of the ReDocRED training set reveals that 7% of these relation instances exhibit a maximum cardinality of 2. This is often due to ground truth comprising both full dates (e.g., "*2 July 1931*") and separate year mentions (e.g., "*1931*") as distinct entities. For example, 15% of training instances of $P570$ (*date of death*) present $k_r > 1$. Consequently, to avoid overly restrictive criteria, we relax the maximum cardinality of such relations to $k_r = 2$. The complete mapping of relations to their respective maximum cardinalities is provided in Table 11 of Appendix D.

Definition 3.4 (Cardinality Violation)**.** Given a relation $r \in \mathcal{R}$ with maximum cardinality k_r, let $O(e_s, r)$ be the set of object entities associated with a subject e_s and relation r in the document d: $O(e_s, r) = \{e_o \in E \mid \langle e_s, r, e_o \rangle \in d, s \neq o\}$. A **cardinality violation** occurs for the subject e_s if the number of distinct extracted triples exceeds the defined threshold: $|O(e_s, r)| > k_r$.

In OWL terms, if $k_r = 1$, the relation r is treated as an `owl:FunctionalProperty`; for $k_r > 1$, it is treated as a `qualified cardinality restriction`.

4 Structural Consistency in DocRE

We analyze the structural consistency of the DocRED datasets using the rules defined above. Given DocRED is the reference test collection for the task, we primarily employ the DocRED distant dataset. We perform the same experiments in the distant datasets produced by the two leading denoising strategies, namely UGDRE (Sun et al., 2023) and DOREMI (Menotti et al., 2026). The results on denoised distant datasets are reported in Appendix A. To compare the amount of noise in distantly and manually annotated data, we also consider the DocRED and ReDocRED manual training datasets. In addition, we perform the same analysis on the ReDocRED evaluation datasets (dev and test), which are the revised version of the split DocRED de-

Dataset	Split	Triples	Invalid Triples (δ)	Missing Inverse (σ)	Asymmetric Inc. (γ)	Cardinality Viol. (θ)
DocRED	Train	38,180	602 (1.58%)	2,696 (64.36%)	154 (0.41%)	10 (0.28%)
ReDocRED	Train	85,932	810 (0.94%)	947 (7.89%)	284 (0.34%)	20 (0.40%)
ReDocRED	Dev	17,284	122 (0.71%)	243 (8.92%)	56 (0.33%)	3 (0.32%)
ReDocRED	Test	17,448	141 (0.81%)	242 (9.27%)	82 (0.48%)	6 (0.65%)
DocRED	Train DS	1,505,638	46,211 (3.07%)	85,065 (40.26%)	12,408 (0.84%)	25 (0.01%)

Table 1: Structural consistency analysis of DocRED datasets. Splits "Train", "Dev", and "Test" consider manually annotated datasets, while split "Train DS" reports the quality of the DocRED distant dataset. For each rule, we report the absolute number of errors in the dataset and its percentage.

velopment dataset. For each dataset, we report the number of errors for each structural constraint (Section 4.1). The code to perform the quality assessment on DocRED and ReDocRED datasets is available online. We also perform a statistically grounded qualitative analysis on invalid triples in the DocRED distant dataset to understand the main cause of such errors (Section 4.2).

4.1 Datasets Consistency Assessment

Table 1 reports the structural consistency of the DocRED datasets. If distant supervision introduces systematic structural degradation, we expect higher violations in the DocRED distant split compared to manual datasets. For each structural constraint defined in Section 3, we report both the absolute number of violations and their percentage, computed according to the specific constraint type.

The percentage of invalid triples (δ) is defined as the ratio between the number of structurally invalid triples and the total number of triples in the dataset. The percentage of inverse triples missing in a dataset (σ) is the ratio between the number of missing inverse triples and the triples involving relations with inverses. The percentage of asymmetric inconsistencies in a dataset (γ) is the ratio between such inconsistencies and the triples involving an asymmetric relation. Finally, the percentage of cardinality violations (θ) is the ratio between the subject entity-relation pair (e_s, r) violating $|O(e_s, r)| > k_r$ over all such pairs involving a relation with bounded maximum cardinality.

As expected, DocRED distant presents much more inconsistencies than the manual datasets. The only exception is on the missing inverse relations, where the worst is the DocRED manual training dataset. This is a known issue in the literature and has been corrected in ReDocRED (Huang et al., 2022; Tan et al., 2022b), as demonstrated by the reduction in missing inverse between the DocRED and ReDocRED manual training dataset (from around 64% to less than 8%). Compared to the ReDocRED manual training dataset, the DocRED distant dataset contains three times more invalid triples, nearly five times more missing inverse relations, and twice as many asymmetric inconsistencies. About cardinality violations, DocRED distant is the dataset containing less errors (0.01%), while the other datasets range between 0.28% and 0.65% of violations. This phenomenon can be attributed to the distant supervision methodology employed in constructing the DocRED distant dataset, which aligns Wikipedia documents with Wikidata triples and thereby preserves their cardinalities.

Also ReDocRED datasets exhibit structural violations, with a small but non-negligible proportion of errors in all splits. Invalid triples constitute between 0.71% and 0.94% of the datasets, missing inverse relations remain below 10%, and asymmetric violations range from 0.33% to 0.48%. Cardinality violations exceed those observed in the DocRED manual training and DocRED distant, reaching 0.65% in the ReDocRED test set. Given the presence of structural violations in the evaluation datasets, models trained to generate well-formed triples may exhibit degraded performance.

4.2 Invalid Triples Analysis

To identify the underlying causes of invalid triples in the distant dataset, we conduct a statistically grounded error analysis. We classify each invalid triple into one of two mutually exclusive categories: (i) relation annotation errors, where, given a correctly typed entity pair, the wrong relation type is assigned, and (ii) entity annotation errors, where one or both entities are incorrectly typed, thereby inducing an invalid relation. For example, the triple $\langle \textit{France}(\texttt{LOC}), \texttt{P16}(\textit{capital}), \textit{Paris}(\texttt{PER}) \rangle$ is considered an entity annotation error because "*Paris*" is mistakenly classified as a person.

Building on this setup, we perform a manual annotation step to estimate how much each of the two error sources contributes to the overall pool of invalid triples. We randomly select 400 invalid triples from DocRED distant and manually annotate each one according to the error taxonomy introduced above. Based on these annotations, we estimate the share of invalid triples attributable to entity errors versus relation errors, and we report a 95% credible interval to capture sampling uncertainty. The evaluation protocol uses unbiased sampling and estimation procedures, providing strong guarantees on the robustness of the reported results. A comprehensive account of the statistical error analysis can be found in Appendix B.

The analysis shows that 69% ($\pm 4.5\%$) of invalid triples stem from errors in entity annotation, while the remaining 31% are attributable to incorrect relation annotations. This pronounced disparity demonstrates that entity-related mistakes are the dominant source of invalid triples. As a result, the findings imply that enhancing named entity recognition and employing targeted denoising methods for entity annotations could markedly improve the structural reliability of distant supervision datasets. More broadly, these findings emphasize that entity extraction is a bottleneck for relation extraction. In end-to-end pipelines, inaccuracies in entity recognition propagate to relation prediction, thereby imposing an upper bound on the overall performance of RE systems. Consequently, improvements in relation extraction are inherently capped by the quality of entity annotations. In this work, we do not modify or refine entity extraction; our intervention is limited to enforcing structural consistency over the extracted triples. As a result, the performance gains we observe should be interpreted as conservative estimates: further improvements could be achieved by jointly addressing entity-level errors. Moreover, since entity misannotations are also present in the evaluation data, correcting structurally inconsistent relations may correspond to semantically correct predictions in real-world settings, yet still be penalized under the current benchmark due to pre-existing entity annotation errors.

5 Training Effect

We investigate if training DocRE models with structurally consistent data improves their performance.

5.1 Cleaning the datasets

We construct a structurally cleaned version of the DocRED distant dataset by eliminating triples that violate the ontological constraints defined in Section 3. Since cardinality violations are extremely rare across all splits, especially in DocRED distant (0.01%), we focus our cleaning procedure on invalid triples, missing inverse relations, and asymmetric violations.

Correcting invalid triples would require manual inspection and revision of both entity types and relation labels. Given the large number of invalid triples in the DocRED distant dataset ($46{,}211$), fully correcting them would entail extensive manual effort and substantial annotation cost. Instead, we adopt a conservative and scalable strategy: for each invalid triple, we remove the entity involved in the violation from the training set, thereby eliminating all triples associated with that entity within the document. Under this strategy, removing an entity effectively removes the offending triple(s) that depend on its incorrect typing or relation assignment. This approach is significantly cheaper than manual correction, as it avoids entity re-labeling while ensuring that no structurally invalid triples remain in the cleaned dataset. Overall, this procedure results in the removal of $208{,}665$ triples and 3 documents from DocRED distant.

Missing inverse relations are added automatically, resulting in the addition of $72,609$ triples. Similarly to invalid triples, solving asymmetric inconsistencies requires the manual evaluation of $12,408$ triples. Thus, we remove all entities involved in at least one asymmetric violation from the dataset, resulting in the removal of $31,178$ triples.

As a result, the *corrected* distant dataset contains $1,338,404$ training instance ($-167,234$ than DocRED distant) and $101,870$ documents (-3 than DocRED distant). The code to correct structural inconsistencies in DocRE datasets is available in the online repository. The code is compatible with any dataset in DocRED format. The repository contains the rules specifically developed for DocRED, but the code can also run on user-defined constraints.

5.2 Experimental Setup

To explore the effect of structural violations on training, we leverage two DocRE models with leading performance and open-source implementations: ATLOP (Zhou et al., 2021) and DREEAM (Ma et al., 2023). We chose two transformer-based models because graph-based models cannot scale to the size of distant datasets. ATLOP exploits a BERT encoder to represent entity pairs enriched with localized context pooling. During prediction,

	PLM	Training Data	Invalid (δ)	Inverse (σ)	Asymmetric (γ)
ATLOP	BERT	DocRED DS	110 (1.61%)	396 (43.04%)	38 (0.56%)
		Corrected	**46 (0.66%)**	**65 (5.61%)**	**4 (0.06%)**
	RoBERTa	DocRED DS	110 (1.56%)	407 (45.22%)	42 (0.60%)
		Corrected	**42 (0.59%)**	**66 (5.68%)**	**0 (0.00%)**
DREEAM	BERT	DocRED DS	59 (1.05%)	266 (41.18%)	30 (0.53%)
		Corrected	**21 (0.31%)**	**115 (11.97%)**	**4 (0.06%)**
	RoBERTa	DocRED DS	52 (0.85%)	333 (44.76%)	48 (0.79%)
		Corrected	**23 (0.34%)**	**92 (10.15%)**	**6 (0.09%)**

Table 2: Structural consistency of DocRE models predictions trained on DocRED Distant ("*DocRED DS*") and its structurally-consistent version ("*Corrected*") and evaluated on **ReDocRED test dataset**. For each rule, we report the absolute number of errors in the predictions and its percentage. Column "*Invalid*" refers to invalid triples, "*Inverse*" to missing inverse triples, and "*Asymmetric*" to asymmetric inconsistencies.

adaptive thresholding is introduced to better handle multi-label classification (Zhou et al., 2021). DREEAM is a teacher-student model, whose architecture is based on ATLOP enriched with an evidence-aware distillation mechanism (Ma et al., 2023). We train both models on the DocRED distant dataset and its *corrected* version, using the ReDocRED development dataset for validation. DREEAM follows a teacher–student paradigm, where the student model follows a first step of "*self-training*" and then is fine-tuned on manual data. During self-training, the student model is trained on distant data and weakly supervised signals from the teacher model. Since our focus is on evaluating the quality of DocRED distant, we consider the DREEAM student model right after self-training, exploiting the teacher model trained on the ReDocRED training dataset. We then evaluate each model on the ReDocRED test dataset and exploit their predictions on these documents to assess the number of ill-formed triples. The code to analyze the structural consistency of the predictions is available online. The code is compatible with any DocRE model predictions formatted as DocRED official results and can be used with any user-defined rules.

5.3 Structural Consistency in predictions

Table 2 reports the structural analysis of predictions by ATLOP and DREEAM models, trained on DocRED distant and its *structurally-consistent* version, all evaluated on ReDocRED test set.

Training DocRE models on structurally inconsistent data results in noisy predictions, with ATLOP appearing slightly more sensitive than DREEAM. The greater robustness of DREEAM may stem from its self-training procedure, which combines distant data with weak supervision signals provided by a teacher model trained on manual data. The proportion of invalid triples in the predictions ranges from 0.85% to 1.61%, about one third of the rate observed in the training data. Missing inverse triples in the training set propagate almost linearly to the predictions: while 40.6% of inverse triples are missing in the dataset, the corresponding figures in the predictions range from 41.18% to 45.22%. A similar pattern holds for asymmetric violations, where the 0.84% rate observed in the training data closely matches the $0.53\% - 0.79\%$ range found in the predictions.

In contrast, training DocRE models without logical inconsistencies consistently improves their ability to generate well-formed triples. Also in this case, the improvement is more pronounced for ATLOP than for DREEAM. On average, the proportion of invalid triples is reduced by -62.92%. The most substantial improvement is observed in inverse relation prediction: for ATLOP-RoBERTa, the percentage of missing inverse relations drops by -87.44%, and for ATLOP-BERT by -86.97%. DREEAM also benefits, with a reduction of -74.13%. In both architectures, asymmetric violations are nearly eliminated; Training ATLOP-RoBERTa on structurally consistent data reduces asymmetric violations from 0.60% to 0.00%.

5.4 Effect on model performance

Having established that structural correction reduces ill-formed predictions, we next examine whether these structural improvements translate into better generalization performance. Table 3 reports the results of DocRE models trained on DocRED distant (row "DocRED DS") and its *structurally-consistent* version (row "Corrected"). We investigate the impact of each noise source in Appendix C. We recall that the corrected dataset contains $167,234$ fewer triples than DocRED distant. To ensure that any performance differences are not merely due to this reduction in training data, we additionally trained the models on a randomly subsampled variant obtained by removing the same number of triples $(167,234)$ from DocRED distant. This random subsampling was repeated with ten different seeds, and the mean performance is reported in the "Random" row. To avoid clutter, Table 3 does not report the standard deviation, as it is always below 1%. All models were validated on the ReDocRED development set and evaluated on

	PLM	Training Data	Precision	IgnPrec	Recall	F1	IgnF1
ATLOP	BERT	DocRED DS	0.7503	0.5877	0.2945	0.4230	0.3924
		Corrected	**0.7665** (+2.16%)	**0.6165** (+4.90%)	**0.3070** (+4.24%)	**0.4384** (+3.64%)	**0.4099** (+4.46%)
		Random	0.7528 (+0.33%)	0.5973 (+1.63%)	0.2964 (+0.65%)	0.4253 (+0.43%)	0.3962 (+0.97%)
	RoBERTa	DocRED DS	0.7571	0.6006	0.3060	0.4358	0.4054
		Corrected	**0.7657** (+1.14%)	**0.6172** (+2.76%)	**0.3120** (+1.96%)	**0.4434** (+1.74%)	**0.4245** (+4.71%)
		Random	0.7578 (+0.09%)	0.6082 (+1.27%)	0.3049 (-0.36%)	0.4348 (-0.23%)	0.4061 (+0.17%)
DREEAM	BERT	DocRED DS	**0.8087**	0.6791	0.2609	0.3945	0.3770
		Corrected	0.8042 (-0.56%)	**0.6941** (+2.21%)	**0.3113** (+19.32%)	**0.4488** (+13.76%)	**0.4298** (+14.01%)
		Random	0.7974 (-1.40%)	0.6746 (-0.66%)	0.2683 (+2.84%)	0.4013 (+1.72%)	0.3837 (+1.78%)
	RoBERTa	DocRED DS	**0.8170**	0.6964	0.2861	0.4238	0.4056
		Corrected	0.8143 (-0.33%)	**0.7083** (+1.71%)	**0.3113** (+8.81%)	**0.4504** (+6.28%)	**0.4353** (+7.32%)
		Random	0.8087 (-1.02%)	0.6974 (+0.14%)	0.2897 (+1.26%)	0.4263 (+0.59%)	0.4090 (+0.84%)

Table 3: Models trained on DocRED distant ("DocRED DS") and its structurally-consistent version ("Corrected") and evaluated on the **ReDocRED test dataset**. Random removal ("Random") was performed ten times with different random seed and the mean performance is reported. The standard deviation is always below 1%. The percentage performance improvement compared to DocRED distant is reported between parentheses.

the ReDocRED test set. Together with precision, recall, and F1 scores, we also ignored metrics, that measures the precision (IGNPREC) or F1 (IGNF1) excluding the entity pairs seen during training.

Training ATLOP on the corrected distant dataset improves performance across all metrics and both transformer backbones. Relative to DocRED distant, structural correction yields average gains of $+1.65\%$ in precision, $+3.10\%$ in recall, and $+2.69\%$ in F1. The improvements are even more pronounced under ignored metrics, where entity pairs seen during training are excluded. On average, IGNPREC increases by $+3.83$, and IGNF1 by $+4.59\%$.

DREEAM exhibits a similar overall trend, with improvements in all metrics except precision. Although precision decreases slightly (by -0.45% on average), this drop is compensated by major gains in the other metrics. Recall improves by $+14.07\%$ on average, with DREEAM-BERT achieving a boost of $+19.32\%$, while F1 rises by $+10.02\%$ on average. Under the ignored setting, IGNPREC increases by $+1.96\%$ and IGNF1 by $+10.67\%$.

Random removal of triples leads to only marginal improvements in most metrics and even degrades performance in several cases: recall and F1 for ATLOP-RoBERTa, precision for both DREEAM variants, and IGNPREC for DREEAM-BERT. These findings indicate that the observed performance gains stem from the correction of structural errors rather than from the mere reduction in the number of training triples or documents.

6 Conclusions

We introduced a framework for assessing the structural consistency of DocRE datasets and analyzed the impact of structural noise on model predictions and downstream performance. We formalized the DocRE task as the construction of a KG under the RDF model and defined a set of logical constraints in OWL to detect inconsistencies. In particular, we identified invalid triples through domain and range constraints on object properties; missing inverse relations via `owl:inverseOf`; asymmetric violations using `owl:AsymmetricProperty`; and relation cardinality violations through `qualified cardinality restrictions`.

Our analysis shows that, compared to the ReDocRED manual training set, the DocRED distant dataset contains three times more invalid triples, nearly five times more missing inverse relations, and twice as many asymmetric inconsistencies. To better understand the cause of invalid triples, we conducted a qualitative analysis on a representative sample. Manual annotation revealed that almost 70% of invalid triples stem from named entity misannotations. This finding suggests that effective denoising strategies should not only target relation labels but also explicitly address entity-level errors.

Finally, we examined whether removing structural inconsistencies from the training data reduces ill-formed predictions and improves model performance. Training DocRE models on structurally consistent data systematically lowers the number of structural violations in the predictions and leads to improved performance across nearly all metrics, with an average gain of $+6.36\%$ in F1 and $+7.62\%$ in IGNF1. These results highlight the importance of structural data quality for both the reliability and the generalization ability of DocRE systems.

7 Limitations

To evaluate the performance gains resulting from structural noise removal, we rely on the ReDocRED test set, which itself contains (a limited number of) logical inconsistencies. An ideal evaluation benchmark for isolating the effects of ill-formed training triples would be entirely free of structural noise. However, re-annotating the ReDocRED test set could raise concerns about annotation bias, as such modifications might be perceived as being tailored to favor our approach.

Our noise mitigation strategy prioritizes entity removal rather than direct error correction. Correcting all ill-formed triples would require extensive manual revision of $58,619$ instances (including invalid triples and asymmetric inconsistencies), which would be prohibitively expensive in terms of time and resources. Instead, we chose to discard the affected entities, ensuring structural consistency while avoiding large-scale manual intervention. Future work could explore systematic re-annotation efforts to preserve a greater portion of the data while further improving dataset quality.

ACTION POINT: We will revise the limitations sections to address potential issues with the quality of the ontology definitions regarding incompleteness, noise, and over-generalizations.

We focus our analysis on the DocRED dataset (Yao et al., 2019) and its different versions (Tan et al., 2022b; Sun et al., 2023; Menotti et al., 2026). We analyzed several DocRE datasets from different domains (especially biomedical), but none of them was suitable for the study. BioRED (Luo et al., 2022) and BC5CDR (Li et al., 2016) are two well-known biomedical DocRE datasets, but they only include one relation type (“`associated_with`”) between different entity types. While GutBrainIE (Martinelli et al., 2026) includes a sufficient variety of entity types and relations, its preprocessing to remove structural inconsistencies before release renders our analysis infeasible.

Finally, our analysis of training effects is restricted to two transformer-based DocRE architectures, ATLOP and DREEAM. Although these models represent strong and widely adopted baselines, future research could extend this investigation to graph-based and generative approaches. To support such efforts, we release our code and the full set of structural rules as open-source resources, enabling the community to reproduce and expand our structural consistency analysis across a broader range of DocRE systems.

On the sample size for the qualitative error analysis: nS=400 yields a 95% HPD interval of ±4.5% around the 0.69 entity-error proportion (Appendix B). Halving the interval width requires $nS \approx 1600$, a four-fold annotation cost without altering the qualitative dominance of entity errors. ACTION POINT: We will make this trade-off explicit in the limitations.

8 Ethical Considerations

Based on our methodology, we do not anticipate any significant ethical concerns. We list some potential ethical aspects and risks of our analysis:

- **Licenses and Data Privacy**: The dataset and models used in this work are open-sourced to minimize the risk of privacy leakage and ensure transparency and accessibility. The DocRE models used in the empirical evaluation are trained and evaluated exclusively on MIT licensed datasets, namely DocRED, ReDocRED, UGDRE, and DOREMI.

- **Hidden Biases**: The leading DocRE models exploited for empirical analysis (ATLOP and DREEAM) are transformer-based models, relying on pre-trained language models such as BERT and RoBERTa. We recognize that pre-trained language models may include hidden biases from their training data, potentially embedding subtle human prejudices. While they perform well in detecting relations, special caution is needed for sensitive relation types or entities, where such biases may become more evident.

- **The use of AI Assistants**: ChatGPT, Perplexity, Writefull, and Grammarly were used purely with the language of the paper to help improve clarity. All scientific content was developed by the authors, and the manuscript has been carefully reviewed and approved in its entirety by all authors.

References

George E. P. Box and George C. Tiao. 2011. *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, New York, NY, USA.

Lawrence D. Brown, Tony T. Cai, and Anirban DasGupta. 2001. Interval Estimation for a Binomial Proportion. *Statistical Science*, 16(2):101–117.

William G. Cochran. 1977. *Sampling Techniques, 3rd Edition*. John Wiley & Sons, New York, NY, USA.

Xin Luna Dong. 2023. Generations of knowledge graphs: The crazy ideas and the business impact. *Proceedings of the VLDB Endowment*, 16(12):4130–4137.

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.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

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:baw068.

Tianming Liang, Yang Liu, Xiaoyan Liu, Hao Zhang, Gaurav Sharma, and Maozu Guo. 2023. Distantly-supervised long-tailed relation extraction using constraint graphs. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):6852–6865.

Yang Liu, Kang Liu, Liheng Xu, and Jun Zhao. 2014. Exploring fine-grained entity type constraints for distantly supervised relation extraction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2107–2116.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding Attention with Evidence for Improving Document-Level Relation Extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983.

Stefano Marchesin and Gianmaria Silvello. 2025. Credible Intervals for Knowledge Graph Accuracy Estimation. *Proceedings of the ACM on Management of Data (SIGMOD)*, 3(3).

Marco Martinelli, Stefano Marchesin, Vanessa Bonato, Giorgio Di Nunzio, Nicola Ferro, Ornella Irrera, Laura Menotti, Federica Vezzani, and Gianmaria Silvello. 2026. A domain-specific curated benchmark for entity and document-level relation extraction. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5693–5711, Rabat, Morocco. Association for Computational Linguistics.

Laura Menotti, Stefano Marchesin, and Gianmaria Silvello. 2026. DOREMI: Optimizing long tail predictions in document-level relation extraction. *Knowledge-Based Systems*, 337:115359.

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

Richard D. Morey, Rink Hoekstra, Jeffrey N. Rouder, Michael D. Lee, and Eric-Jan Wagenmakers. 2016. The Fallacy of Placing Confidence in Confidence Intervals. *Psychonomic Bulletin & Review*, 23:103–123.

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 *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1239–1250.

Qi Sun, Kun Huang, Xiaocui Yang, Pengfei Hong, Kun Zhang, and Soujanya Poria. 2023. Uncertainty Guided Label Denoising for Document-level Distant Relation Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15960–15973.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting DocRED - Addressing the False Negative Problem in Relation Extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487.

Chaojun Xiao, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Maosong Sun, Fen Lin, and Leyu Lin. 2020. Denoising Relation Extraction from Document-level Distant Supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3683–3688.

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.

Fu Zhang, Qi Miao, Jingwei Cheng, Hongsen Yu, Yi Yan, Xin Li, and Yongxue Wu. 2024. SRF: Enhancing Document-Level Relation Extraction with a Novel Secondary Reasoning Framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15426–15439.

Fu Zhang, Yi Yan, and Jingwei Cheng. 2025. Document-level relation extraction with global relations and entity pair reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19556–19567. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14612–14620.

A Denoised Datasets

We perform the same structural consistency analysis presented for DocRED distant on four denoised versions of the distant dataset, generated by two top-performing denoising techniques. UGDRE is a label denoising strategy that exploits Monte Carlo dropout to estimate pseudo-label uncertainty and filter out low-confidence triples produced by a DocRE model trained on the manual training data (Sun et al., 2023). Given the manual DocRED training has been revised to produce ReDocRED, UGDRE generates two denoised distant datasets: one exploiting the DocRED manual as the training dataset (UGDRE-ReDocRED) and the other using the ReDocRED training dataset (UGDRE-ReDocRED). The second denoising strategy is DOREMI, an active learning system tailored for enhancing long-tail triples (Menotti et al., 2026). Like UGDRE, DOREMI exploits a set of DocRE models pre-trained on a manual training dataset. Thus, this approach also generates two denoised versions of DocRED distant: DOREMI-DocRED and DOREMI-ReDocRED.

A.1 Consistency Assessment

Table 4 reports the structural consistency of the four distant datasets, produced by performing the same syntactic analysis described in Section 4.1.

The percentage of invalid triples is almost doubled in UGDRE-based datasets than DOREMI. In particular, UGDRE-DocRED has almost the same percentage of invalid triples of the DocRED distant (3.01% vs 3.07%), hinting at the inability of current denoising techniques to eliminate structural inconsistencies. About missing inverse relations, the denoised datasets generated by using the DocRED training dataset misses much more inverse triples than the two datasets that uses ReDocRED. Indeed, the percentage of missing inverse relations in UGDRE-RedocRED is half of UGDRE-DocRED (25.42% vs 53.50%) while it is three times smaller between DOREMI-ReDocRED and DOREMI-DocRED (25.42% vs 75.13%). This is expected given that ReDocRED solves the false negative issues of DocRED (Tan et al., 2022b). Asymmetric inconsistencies are more evident in UGDRE datasets than the DOREMI ones. For instance, the percentage of asymmetric inconsistencies of UGDRE-DocRED is double than DOREMI-DocRED (0.77% vs 0.36%). Nevertheless, all denoised datasets reports a smaller percentage of

Dataset	Instances	Invalid Triples (δ)	Missing Inverse (σ)	Asymmetric Inc. (γ)	Cardinality Viol. (θ)
UGDRE-DocRED	1,717,161	51,671 (3.01%)	112,912 (53.50%)	13,026 (0.77%)	74 (0.03%)
DOREMI-DocRED	1,704,161	29,799 (1.75%)	128,064 (75.13%)	6,030 (0.36%)	1,179 (0.73%)
UGDRE-ReDocRED	3,379,590	94,290 (2.79%)	122,913 (25.42%)	19,416 (0.59%)	1,426 (0.52%)
DOREMI-ReDocRED	3,957,238	69,035 (1.74%)	111,089 (22.00%)	15,244 (0.39%)	3,428 (1.43%)

Table 4: Structural consistency analysis of the denoised distant datasets. For each rule, we report the absolute number of errors in the dataset and its percentage.

asymmetric inconsistencies than DocRED distant (i.e., 0.84%). The percentage of cardinality violations are much larger than DocRED distant, which was almost zero (0.01%). DOREMI datasets shows the highest number of errors, peaking at 1.43% for DOREMI-ReDocRED. In general, for both denoising strategies, the DocRED-based denoised datasets have way less cardinality violations than the ReDocRED-based. In particular, UGDRE-DocRED has almost seventeen times less violations than UGDRE-ReDocRED (0.03% vs 0.52%), while DOREMI-DocRED has half violations than DOREMI-ReDocRED (0.73% vs 1.43%). This behavior may be attributed to the different dataset construction technique with respect to DocRED distant and to the fact that ReDocRED training dataset has much more training triples than the DocRED manual training dataset ($85,932$ vs $38,180$), training the model to predict more positive triples.

Overall, state-of-the-art denoising techniques fail to mitigate structural inconsistencies and, in the case of cardinality violations, even amplify them.

A.2 Training effect

This section describes the structural consistency of the predictions of ATLOP and DREEAM trained on the four denoised datasets and their *corrected* versions and their performance. The correction of the datasets follows the procedure described in Section 5.1 and the adopted experimental setup is described in Section 5.2.

A.2.1 Models Predictions

Table 5 reports the structural consistency of the predictions of ATLOP and DREEAM trained on the four denoised datasets and their *corrected* versions. As for DocRED, training without structural noise systematically lowers the presence of ill-formatted triples in the predictions.

The percentage of missing inverse triples exhibits the most evident improvements, with corrected DOREMI-DocRED managing to reduce missing inverse relations by -86.45% on average in both DocRE models and configurations. The improvements in ReDocRED-based datasets for missing inverse triples are more contained since the percentage of errors in the datasets was already much lower than DocRED-based datasets, but still ranging from -55.01% to -83.22% percentage points.

The improvements in invalid triples and asym-

Model	Training Data	Invalid (δ)	Inverse (σ)	Asymmetric (γ)
ATLOP BERT	UGDRE-Doc	106 (1.36%)	528 (59.26%)	38 (0.49%)
	Corrected	**42 (0.54%)**	**65 (5.40%)**	**2 (0.03%)**
	DOREMI-Doc	47 (0.72%)	335 (79.01%)	26 (0.40%)
	Corrected	**33 (0.46%)**	**90 (10.79%)**	**10 (0.14%)**
	UGDRE-ReD	188 (1.24%)	579 (26.82%)	66 (0.44%)
	Corrected	**118 (0.79%)**	**149 (6.16%)**	**6 (0.04%)**
	DOREMI-ReD	164 (1.02%)	381 (19.44%)	50 (0.31%)
	Corrected	**144 (0.96%)**	**158 (8.27%)**	**14 (0.09%)**
ATLOP RoBERTa	UGDRE-Doc	110 (1.41%)	505 (56.11%)	42 (0.54%)
	Corrected	**39 (0.48%)**	**65 (5.15%)**	**0 (0.00%)**
	DOREMI-Doc	48 (0.74%)	365 (74.95%)	16 (0.25%)
	Corrected	**37 (0.53%)**	**93 (12.35%)**	**10 (0.14%)**
	UGDRE-ReD	197 (1.29%)	621 (28.19%)	62 (0.41%)
	Corrected	**99 (0.64%)**	**121 (4.73%)**	**4 (0.03%)**
	DOREMI-ReD	194 (1.19%)	407 (20.65%)	60 (0.37%)
	Corrected	**120 (0.79%)**	**181 (9.29%)**	**12 (0.08%)**
DREEAM BERT	UGDRE-Doc	106 (1.36%)	528 (59.26%)	38 (0.49%)
	Corrected	**27 (0.39%)**	**71 (7.63%)**	**4 (0.06%)**
	DOREMI-Doc	69 (1.04%)	363 (69.67%)	26 (0.40%)
	Corrected	**32 (0.46%)**	**58 (7.81%)**	**10 (0.14%)**
	UGDRE-ReD	188 (1.24%)	579 (26.82%)	66 (0.44%)
	Corrected	**87 (0.59%)**	**158 (6.55%)**	**12 (0.08%)**
	DOREMI-ReD	135 (0.85%)	287 (15.41%)	42 (0.27%)
	Corrected	**119 (0.73%)**	**142 (6.59%)**	**12 (0.07%)**
DREEAM RoBERTa	UGDRE-Doc	115 (1.37%)	576 (56.92%)	110 (1.32%)
	Corrected	**27 (0.36%)**	**86 (7.96%)**	**10 (0.13%)**
	DOREMI-Doc	54 (0.71%)	461 (76.71%)	36 (0.48%)
	Corrected	**52 (0.68%)**	**111 (9.85%)**	**24 (0.32%)**
	UGDRE-ReD	116 (0.78%)	451 (22.09%)	76 (0.51%)
	Corrected	**73 (0.48%)**	**129 (5.16%)**	**30 (0.20%)**
	DOREMI-ReD	103 (0.64%)	306 (16.59%)	48 (0.30%)
	Corrected	**100 (0.61%)**	**118 (5.45%)**	**18 (0.11%)**

Table 5: Structural consistency of DocRE models predictions. The trained on different denoised distant datasets and evaluated on the **ReDocRED test dataset**. For space reasons, we use suffix "-Doc" as a diminutive for DocRED and "-ReD" as a diminutive for ReDocRED. For each rule, we report the absolute number of errors in the predictions and its percentage. Column "*Invalid*" refers to invalid triples, "*Inverse*" to missing inverse triples, and "*Asymmetric*" to asymmetric inconsistencies.

Model	Training Data	Precision	IgnPrec	Recall	F1	IgnF1
ATLOP BERT	UGDRE-Doc	0.7881	0.6592	0.3512	0.4858	0.4582
	Corrected	**0.8002** (+1.54%)	**0.6791** (+3.02%)	**0.3593** (+2.31%)	**0.4959** (+2.08%)	**0.4700** (+2.58%)
	DOREMI-Doc	**0.8686**	**0.7903**	0.3234	0.4702	0.4589
	Corrected	0.8439 (-2.84%)	0.7716 (-2.37%)	**0.3458** (+6.93%)	**0.4905** (+4.32%)	**0.4776** (+4.07%)
	UGDRE-ReD	0.8044	**0.7445**	**0.7007**	**0.7490**	**0.7220**
	Corrected	**0.8069** (+0.31%)	0.7093 (-4.73%)	0.6937 (-1.00%)	0.7461 (-0.39%)	0.7014 (-2.85%)
	DOREMI-ReD	0.7480	0.6297	**0.6914**	**0.7186**	0.6591
	Corrected	**0.7702** (+2.97%)	**0.6650** (+5.61%)	0.6602 (-4.51%)	0.7110 (-1.06%)	**0.6626** (+0.53%)
ATLOP RoBERTa	UGDRE-Doc	0.7977	0.6761	0.3573	0.4936	0.4676
	Corrected	**0.8012** (+0.44%)	**0.6832** (+1.05%)	**0.3755** (+5.09%)	**0.5114** (+3.61%)	**0.4847** (+3.66%)
	DOREMI-Doc	**0.8701**	**0.8031**	0.3228	0.4709	0.4605
	Corrected	0.8605 (-1.10%)	0.7942 (-1.11%)	**0.3439** (+6.54%)	**0.4914** (+4.35%)	**0.4800** (+4.23%)
	UGDRE-ReD	**0.8090**	**0.7502**	0.7087	**0.7555**	**0.7288**
	Corrected	0.8025 (-0.80%)	0.7048 (-6.05%)	**0.7129** (+0.59%)	0.7551 (-0.05%)	0.7088 (-2.74%)
	DOREMI-ReD	0.7486	0.6311	**0.7023**	**0.7247**	0.6648
	Corrected	**0.7665** (+2.39%)	**0.6624** (+4.96%)	0.6704 (-4.54%)	0.7152 (-1.31%)	**0.6664** (+0.24%)
DREEAM BERT	UGDRE-Doc	0.8483	0.7507	0.3247	0.4696	0.4533
	Corrected	**0.8562** (+0.93%)	**0.7680** (+2.30%)	**0.3413** (+5.12%)	**0.4881** (+3.93%)	**0.4726** (+4.25%)
	DOREMI-Doc	0.8796	0.8176	0.3334	0.4836	0.4737
	Corrected	**0.8853** (+0.64%)	**0.8283** (+1.31%)	**0.3564** (+6.88%)	**0.5082** (+5.08%)	**0.4983** (+5.20%)
	UGDRE-ReD	**0.8265**	**0.7740**	**0.6966**	**0.7560**	**0.7333**
	Corrected	0.8180 (-1.03%)	0.7281 (-5.04%)	0.6958 (-0.12%)	0.7520 (-0.54%)	0.7116 (-2.96%)
	DOREMI-ReD	**0.7701**	**0.6576**	0.7037	**0.7354**	0.6799
	Corrected	0.7617 (-1.09%)	0.6574 (-0.04%)	**0.7083** (+0.65%)	0.7341 (-0.19%)	**0.6819** (+0.29%)
DREEAM RoBERTa	UGDRE-Doc	0.7806	0.6749	**0.3748**	0.5065	0.4820
	Corrected	**0.8575** (+9.85%)	**0.7745** (+14.75%)	0.3692 (-1.51%)	**0.5161** (+1.91%)	**0.5000** (+3.74%)
	DOREMI-Doc	0.8737	0.8131	0.3782	0.5279	0.5163
	Corrected	**0.8815** (+0.89%)	**0.8249** (+1.45%)	**0.3876** (+2.47%)	**0.5384** (+1.99%)	**0.5273** (+2.14%)
	UGDRE-ReD	**0.8464**	**0.7996**	0.7252	**0.7811**	0.7606
	Corrected	0.8329 (-1.58%)	0.7500 (-6.20%)	**0.7321** (+0.96%)	0.7793 (-0.23%)	**0.7410** (-2.58%)
	DOREMI-ReD	**0.7939**	**0.6894**	0.7267	**0.7588**	**0.7075**
	Corrected	0.7785 (-1.95%)	0.6801 (-1.36%)	**0.7334** (+0.91%)	0.7553 (-0.48%)	0.7057 (-0.27%)

Table 6: ATLOP and DREEAM trained on different denoised distant datasets and evaluated on the **ReDocRED test dataset**. For space reasons, we use suffix "-Doc" as a diminutive for DocRED and "-ReD" as a diminutive for ReDocRED.

metric inconsistencies are similar to those reported for DocRED. The highest improvement for invalid triples is in UGDRE-DocRED for DREEAM-RoBERTa, where the percentage of errors drops from 1.37% to 0.36%, registering a relative reduction of -73.72%. The same configuration and dataset shows the biggest reduction in asymmetric inconsistencies, going from 1.32% to 0.13% (-90.15% relative decrease). In addition, training ATLOP-RoBERTa with the corrected version of UGDRE-DocRED completely eliminates the presence of asymmetric inconsistencies in the predictions.

A.2.2 Models Performance

Table 6 reports the performance of ATLOP and DREEAM trained on the different denoised datasets and their corrected versions evaluated on the ReDocRED test dataset. While the correction of the DocRED distant supervision dataset produced systematic performance gains across both models and configurations, the denoised datasets did not always benefit from it. However, such a behavior may be attributed to the evaluation dataset containing some ill-formed triples (see Section 4). Overall, there is a performance improvement in at least one metric for all models and configurations, except for DREEAM-BERT trained on the corrected version of UGDRE-ReDocRED. In 47 out of 80 cases (59%), our approach yields performance improvements, demonstrating a net positive effect overall. Correcting the DocRED-based denoised datasets exhibit a performance improvement in more cases than ReDocRED-based datasets. Considering DocRED-based datasets, precision and ignored precision (IGNPREC) shows an improvement in six cases out of eight – with a peak of $+9.85\%$ and $+14.75\%$ respectively for DREEAM-RoBERTa trained on the corrected UGDRE-DocRED – recall in seven, while F1 and ignored F1 (IGNF1) increase for all models and configurations.

B Statistical Error Analysis

To investigate the causes of invalid triples, we adopt a statistically grounded evaluation procedure. Recall that errors only arise from two mutually exclusive types: (i) errors originating from incorrect relation annotations, and (ii) errors caused by incorrect entity annotations. This allows us to estimate the relative proportion of these two error sources within the population of invalid triples. However, since annotating all invalid triples within DocRED distant would be infeasible in terms of time and cost, we instead rely on binomial proportion estimation techniques (Brown et al., 2001). This approach enables us to obtain statistically reliable estimates while limiting the amount of manual annotation required.

Specifically, we draw a sample using Simple Random Sampling (SRS) of size $n_{\mathcal{S}} = 400$ from the DocRED distant dataset population of invalid triples, ensuring representativeness of the sample (Cochran, 1977). Each sampled triple is manually annotated according to the error classification described above. We formalize the annotation process as a binomial indicator function $\mathbb{1}(\langle e_s, r, e_o \rangle)$, which takes value 1 if the triple contains an entity annotation error (e.g., incorrect entity type) and 0 if it contains a relation annotation error.

After annotation, we compute the sample proportion $\hat{\mu} = \frac{\sum_{\langle e_s, r, e_o \rangle \in \mathcal{S}} \mathbb{1}(\langle e_s, r, e_o \rangle)}{n_{\mathcal{S}}}$, which is an unbiased estimator of the true population proportion (Cochran, 1977). To quantify sampling uncertainty, we accompany this estimate with a $1 - \alpha$ credible interval. We employ Highest Posterior Density (HPD) intervals (Box and Tiao, 2011), which are well suited for uncertainty quantification in KG evaluation settings (Marchesin and Silvello, 2025). Unlike (frequentist) confidence intervals, HPD intervals operate directly in the probabilistic space and avoid common interpretational pitfalls (Morey et al., 2016), making them appropriate for one-shot evaluation scenarios (Marchesin and Silvello, 2025). We set $\alpha = 0.05$, yielding a 95% credible interval.

C Individual Structural Constraint Impact

We performed a study on the impact of each noise source to decide which structural inconsistencies to include in the study. In particular, we trained ATLOP on different versions of the DocRED distant dataset, where we corrected a single structural inconsistency (invalid triples, missing inverse, or asymmetric inconsistency). Table 7 reports the micro-averaged performance of ATLOP and DREEAM trained on the different versions of the DocRED distant dataset evaluated on the development dataset. We report the performance of the original distant dataset (row "DocRED DS"). We also report the performance when cleaning only a single constraints. For instance, row "w/o invalid" reports the performance of ATLOP and DREEAM trained on the DocRED distant dataset cleaned from invalid triples only.

Correcting individual sources of noise yields consistent performance gains for ATLOP in both transformer configurations, with the magnitude of improvement varying by noise type. Invalid triples represent the predominant source of noise; for instance, in ATLOP-BERT, correcting invalid triples of the DocRED distant dataset results in a precision improvement of $+0.35\%$, recall $+0.87\%$, and F1 $+0.73\%$. The gain is even more pronounced in ignored metrics (IGNPREC $+2.08\%$ and IGNF1 $+1.27\%$), when entity pairs seen during training are discarded. Addressing the missing inverse triples results in the highest recall increase ($+3.55\%$ in ATLOP-BERT), but shows a slight decrease in terms of precision (-0.68% in ATLOP-BERT and -0.11% in ATLOP-RoBERTa). Integrating all corrections cumulatively achieves the highest performance gains (precision $+0.60\%$, IGNPREC $+2.11\%$, recall $+4.05\%$, F1 $+3.41\%$, IGNF1 $+3.41\%$ in ATLOP-BERT). For ATLOP-RoBERTa, correcting all structural inconsistencies registers the highest average improvement across all metrics. The configuration shows the best precision improvement (precision $+2.66\%$, IGNPREC $+5.77\%$) and the second best gain in recall ($+2.85\%$), F1 ($+2.80\%$), and IGNF1 ($+3.82\%$).

	PLM	Training Data	Precision	IgnPrec	Recall	F1	IgnF1
ATLOP	BERT	DocRED DS	0.7657	0.6069	0.2983	0.4293	0.4000
		w/o invalid	0.7684 (+0.35%)	0.6196 (+2.08%)	0.3009 (+0.87%)	0.4324 (+0.73%)	0.4050 (+1.27%)
		w/o inverse	0.7605 (-0.68%)	0.5980 (-1.46%)	0.3088 (+3.55%)	0.4393 (+2.33%)	0.4073 (+1.84%)
		w/o asymmetric	0.7660 (+0.03%)	0.6094 (+0.41%)	0.3009 (+0.89%)	0.4321 (+0.65%)	0.4029 (+0.73%)
		All rules	**0.7704** (+0.60%)	**0.6197** (+2.11%)	**0.3103** (+4.05%)	**0.4424** (+3.06%)	**0.4136** (+3.41%)
	RoBERTa	DocRED DS	0.7603	0.6041	0.3106	0.4410	0.4102
		w/o invalid	0.7623 (+0.26%)	0.6267 (+3.74%)	0.3184 (+2.53%)	0.4492 (+1.86%)	0.4223 (+2.94%)
		w/o inverse	0.7595 (-0.11%)	0.6061 (+0.32%)	**0.3305** (+6.43%)	**0.4606** (+4.44%)	**0.4277** (+4.27%)
		w/o asymmetric	0.7648 (+0.59%)	0.6093 (+0.85%)	0.3105 (-0.04%)	0.4416 (+0.14%)	0.4113 (+0.26%)
		All rules	**0.7806** (+2.66%)	**0.6390** (+5.77%)	0.3194 (+2.85%)	0.4533 (+2.80%)	0.4259 (+3.82%)
DREEAM	BERT	DocRED DS	0.8116	**0.7667**	0.2613	0.3953	0.3897
		w/o invalid	0.7970 (-1.80%)	0.6847 (-10.69%)	0.2747 (+5.14%)	0.4086 (+3.36%)	0.3921 (+0.61%)
		w/o inverse	0.8109 (-0.09%)	0.6839 (-10.80%)	0.2772 (+6.09%)	0.4132 (+4.52%)	0.3945 (+1.22%)
		w/o asymmetric	**0.8148** (+0.39%)	0.6888 (-10.16%)	0.2635 (+0.86%)	0.3983 (+0.75%)	0.3812 (-2.19%)
		All rules	0.7992 (-1.54%)	0.6909 (-9.89%)	**0.3018** (+15.52%)	**0.4382** (+10.85%)	**0.4201** (+7.80%)
	RoBERTa	DocRED DS	0.8270	**0.7885**	0.2864	0.4255	0.4202
		w/o invalid	0.8220 (-0.60%)	0.7184 (-8.90%)	0.2950 (+2.99%)	0.4342 (+2.04%)	0.4183 (-0.47%)
		w/o inverse	0.8252 (-0.21%)	0.7102 (-9.93%)	0.2937 (+2.52%)	0.4332 (+1.81%)	0.4155 (-1.12%)
		w/o asymmetric	**0.8295** (+0.31%)	0.7158 (-9.23%)	0.2787 (-2.71%)	0.4172 (-1.95%)	0.4012 (-4.53%)
		All rules	0.8213 (-0.68%)	0.7179 (-8.96%)	**0.3066** (+7.05%)	**0.4466** (+4.95%)	**0.4297** (+2.26%)

Table 7: Models trained on DocRED distant (“DocRED DS”) and its structurally-consistent versions and evaluated on the **ReDocRED dev dataset**. Row “All rules" considers the dataset cleaned from all structural constraints, row “w/o invalid" considers the dataset dataset cleaned from invalid triples only, row “w/o inverse" dataset cleaned from missing inverse triples only, and row “w/o asymmetric" dataset cleaned from asymmetric inconsistencies only. The percentage performance improvement compared to DocRED distant is reported between parentheses.

Correcting different types of rules for DREEAM shows the same performance trend as in ATLOP; invalid rules are the most influential source of noise, and inverse rules boost the recall. Applying all structural constraints together achieves the best results, showing a significant performance improvement in recall ($+15.52\%$ in DREEAM-BERT, $+7.05\%$ for DREEAM-RoBERTa) and F1 ($+10.85\%$ in DREEAM-BERT, $+4.95\%$ for DREEAM-RoBERTa), while registering a small precision degradation (-1.54% in DREEAM-BERT, -0.68% for DREEAM-RoBERTa).

Both DocRE models and both configurations demonstrate that each source of noise has a significant impact on the model performance: removing invalid triples improves precision, while adding missing inverse relations increases recall. However, integrating all structural constraints together harnesses the strength of each individual rule, producing the best overall performance.

D DocRED Structural Constraints

This section reports the structural constraints exploited for our study. The rules are also available in JSON format in Github. [5] Tables 8 and 9 report the domain and range entity type allowed for each DocRED relation. Table 10 reports the relations with an inverse in DocRED. To conclude, Table 11 reports the relations with an upper bound on the maximum cardinality.

[5] Folder “data/rules/" at: https://anonymous.4open.science/r/DocRE-StructuralConsistency-D01F/

<table>
<thead>
<tr><th>Wikidata ID</th><th>Name</th><th>Domain</th><th>Range</th></tr>
</thead>
<tbody>
<tr><td>P6</td><td>head of government</td><td>LOC|ORG|MISC</td><td>PER|MISC</td></tr>
<tr><td>P17</td><td>country</td><td>ORG|LOC|PER|MISC|TIME</td><td>LOC|ORG|MISC</td></tr>
<tr><td>P19</td><td>place of birth</td><td>PER|MISC</td><td>LOC|MISC</td></tr>
<tr><td>P20</td><td>place of death</td><td>PER|MISC</td><td>LOC|MISC</td></tr>
<tr><td>P22</td><td>father</td><td>PER|MISC</td><td>PER|MISC</td></tr>
<tr><td>P25</td><td>mother</td><td>PER|MISC</td><td>PER|MISC</td></tr>
<tr><td>P26</td><td>spouse</td><td>PER|MISC</td><td>PER|MISC</td></tr>
<tr><td>P27</td><td>country of citizenship</td><td>PER|ORG|MISC</td><td>LOC|ORG|MISC</td></tr>
<tr><td>P30</td><td>continent</td><td>ORG|LOC|MISC</td><td>LOC|MISC</td></tr>
<tr><td>P31</td><td>instance of</td><td>PER|ORG|LOC|TIME|NUM|MISC</td><td>PER|ORG|LOC|TIME|NUM|MISC</td></tr>
<tr><td>P35</td><td>head of state</td><td>LOC|MISC</td><td>PER|MISC</td></tr>
<tr><td>P36</td><td>capital</td><td>LOC|ORG|MISC</td><td>LOC|MISC</td></tr>
<tr><td>P37</td><td>official language</td><td>LOC|ORG|MISC</td><td>MISC</td></tr>
<tr><td>P39</td><td>position held</td><td>PER|MISC</td><td>ORG|MISC</td></tr>
<tr><td>P40</td><td>child</td><td>PER|MISC</td><td>PER|MISC</td></tr>
<tr><td>P50</td><td>author</td><td>MISC</td><td>PER|ORG|MISC</td></tr>
<tr><td>P54</td><td>member of sports team</td><td>PER|NUM|MISC</td><td>ORG|LOC|MISC</td></tr>
<tr><td>P57</td><td>director</td><td>MISC</td><td>PER|MISC</td></tr>
<tr><td>P58</td><td>screenwriter</td><td>MISC</td><td>PER|MISC</td></tr>
<tr><td>P69</td><td>educated at</td><td>PER|MISC</td><td>ORG|LOC|MISC</td></tr>
<tr><td>P86</td><td>composer</td><td>NUM|MISC</td><td>PER|MISC</td></tr>
<tr><td>P102</td><td>member of political party</td><td>PER|MISC</td><td>ORG|MISC</td></tr>
<tr><td>P108</td><td>employer</td><td>PER|MISC</td><td>LOC|ORG|MISC</td></tr>
<tr><td>P112</td><td>founded by</td><td>LOC|ORG|MISC</td><td>PER|ORG|LOC|MISC</td></tr>
<tr><td>P118</td><td>league</td><td>PER|ORG|MISC</td><td>ORG|LOC|MISC</td></tr>
<tr><td>P123</td><td>publisher</td><td>MISC</td><td>PER|ORG|MISC</td></tr>
<tr><td>P127</td><td>owned by</td><td>ORG|LOC|PER|MISC</td><td>PER|ORG|LOC|MISC</td></tr>
<tr><td>P131</td><td>located in the administrative territorial entity</td><td>ORG|LOC|MISC</td><td>LOC|MISC</td></tr>
<tr><td>P136</td><td>genre</td><td>PER|ORG|MISC</td><td>ORG|LOC|MISC</td></tr>
<tr><td>P137</td><td>operator</td><td>LOC|ORG|MISC</td><td>ORG|LOC|PER|MISC</td></tr>
<tr><td>P140</td><td>religion</td><td>PER|ORG|LOC|MISC</td><td>ORG|MISC</td></tr>
<tr><td>P150</td><td>contains administrative territorial entity</td><td>ORG|LOC|MISC</td><td>ORG|LOC|MISC</td></tr>
<tr><td>P155</td><td>follows</td><td>PER|ORG|LOC|TIME|NUM|MISC</td><td>PER|ORG|LOC|TIME|NUM|MISC</td></tr>
<tr><td>P156</td><td>followed by</td><td>PER|ORG|LOC|TIME|NUM|MISC</td><td>PER|ORG|LOC|TIME|NUM|MISC</td></tr>
<tr><td>P159</td><td>headquarters location</td><td>ORG|LOC|MISC</td><td>ORG|LOC|MISC</td></tr>
<tr><td>P161</td><td>cast member</td><td>MISC</td><td>PER|MISC</td></tr>
<tr><td>P162</td><td>producer</td><td>MISC</td><td>PER|ORG|MISC</td></tr>
<tr><td>P166</td><td>award received</td><td>PER|LOC|ORG|MISC</td><td>MISC|ORG</td></tr>
<tr><td>P170</td><td>creator</td><td>LOC|ORG|PER|NUM|MISC</td><td>PER|ORG|MISC</td></tr>
<tr><td>P171</td><td>parent taxon</td><td>PER|ORG|LOC|TIME|NUM|MISC</td><td>PER|ORG|LOC|TIME|NUM|MISC</td></tr>
<tr><td>P172</td><td>ethnic group</td><td>PER|ORG|LOC|MISC</td><td>LOC|ORG|MISC</td></tr>
<tr><td>P175</td><td>performer</td><td>MISC|ORG|PER</td><td>PER|ORG|MISC</td></tr>
<tr><td>P176</td><td>manufacturer</td><td>MISC|ORG</td><td>ORG|PER|MISC</td></tr>
<tr><td>P178</td><td>developer</td><td>MISC</td><td>ORG|PER|MISC</td></tr>
<tr><td>P179</td><td>series</td><td>MISC|NUM</td><td>ORG|MISC</td></tr>
<tr><td>P190</td><td>sister city</td><td>LOC|MISC</td><td>LOC|MISC</td></tr>
<tr><td>P194</td><td>legislative body</td><td>LOC|ORG|MISC</td><td>LOC|ORG|MISC</td></tr>
<tr><td>P205</td><td>basin country</td><td>LOC|MISC</td><td>LOC|MISC</td></tr>
</tbody>
</table>

Table 8: Domain and Range constraints of DocRED relations (I). For each relation, we report the Wikidata ID and a textual description.

Wikidata ID	Name	Domain	Range
P206	located in or next to body of water	LOC|ORG|MISC	LOC|MISC
P241	military branch	PER|ORG|MISC	ORG|MISC
P264	record label	PER|ORG|MISC	ORG|MISC
P272	production company	ORG|MISC	ORG|MISC
P276	location	PER|ORG|LOC|TIME|NUM|MISC	ORG|LOC|MISC
P279	subclass of	PER|ORG|LOC|TIME|NUM|MISC	PER|ORG|LOC|TIME|NUM|MISC
P355	subsidiary	PER|ORG|MISC	LOC|ORG|MISC
P361	part of	PER|ORG|LOC|TIME|NUM|MISC	PER|ORG|LOC|TIME|NUM|MISC
P364	original language of work	MISC	LOC|MISC
P400	platform	MISC	MISC
P403	mouth of the watercourse	LOC|MISC	LOC|MISC
P449	original network	MISC	ORG|MISC
P463	member of	PER|ORG|LOC|MISC	LOC|ORG|MISC
P488	chairperson	ORG|LOC|MISC	PER|MISC
P495	country of origin	MISC	LOC|MISC
P527	has part	PER|ORG|LOC|TIME|NUM|MISC	PER|ORG|LOC|TIME|NUM|MISC
P551	residence	PER|MISC	LOC|MISC
P569	date of birth	PER|MISC	TIME|MISC
P570	date of death	PER|MISC	TIME|MISC
P571	inception	ORG|LOC|TIME|MISC	TIME|MISC
P576	dissolved, abolished or demolished	ORG|LOC|MISC	TIME|MISC
P577	publication date	MISC	TIME|MISC
P580	start time	PER|ORG|LOC|TIME|NUM|MISC	TIME|MISC
P582	end time	PER|ORG|LOC|TIME|NUM|MISC	TIME|MISC
P585	point in time	PER|ORG|LOC|TIME|NUM|MISC	TIME|MISC
P607	conflict	PER|ORG|LOC|MISC	ORG|TIME|LOC|MISC
P674	characters	MISC	PER|ORG|MISC|LOC
P676	lyrics by	MISC	PER|ORG|MISC
P706	located on terrain feature	LOC|ORG|MISC	LOC|MISC
P710	participant	PER|ORG|LOC|MISC	PER|ORG|LOC|MISC
P737	influenced by	PER|ORG|LOC|TIME|NUM|MISC	PER|ORG|LOC|TIME|NUM|MISC
P740	location of formation	ORG|MISC	LOC|MISC
P749	parent organization	LOC|ORG|MISC	PER|ORG|MISC
P800	notable work	PER|ORG|MISC	PER|LOC|ORG|NUM|MISC
P807	separated from	LOC|ORG|MISC	LOC|ORG|MISC
P840	narrative location	MISC	LOC|MISC
P937	work location	PER|ORG|MISC	LOC|MISC
P1001	applies to jurisdiction	PER|ORG|MISC	LOC|ORG|MISC
P1056	product or material produced	PER|ORG|MISC	MISC|ORG
P1198	unemployment rate	LOC|MISC	NUM|MISC
P1336	territory claimed by	LOC|MISC	PER|ORG|LOC|MISC
P1344	participant of	PER|ORG|LOC|MISC	PER|ORG|LOC|MISC
P1365	replaces	PER|ORG|LOC|MISC	PER|ORG|LOC|MISC
P1366	replaced by	PER|ORG|LOC|MISC	PER|ORG|LOC|MISC
P1376	capital of	LOC|MISC	LOC|ORG|MISC
P1412	languages spoken, written or signed	PER|MISC	LOC|MISC
P1441	present in work	PER|ORG|LOC|MISC	MISC
P3373	sibling	PER|MISC	PER|MISC

Table 9: Domain and Range constraints of DocRED relations (II). For each relation, we report the Wikidata ID and a textual description (column “Name”).

Wikidata ID	Name	Inverse ID	Inverse Name
P22	father	P40	child
P25	mother	P40	child
P26	spouse	P26	spouse
P36	capital	P1376	capital of
P50	author	P800	notable work
P57	director	P800	notable work
P86	composer	P800	notable work
P155	follows	P156	followed by
P156	followed by	P155	follows
P170	creator	P800	notable work
P176	manufacturer	P1056	product or material produced
P355	subsidiary	P749	parent organization
P361	part of	P527	has part
P527	has part	P361	part of
P674	characters	P1441	present in work
P710	participant	P1344	participant of
P749	parent organization	P355	subsidiary
P1344	participant of	P710	participant
P1365	replaces	P1366	replaced by
P1366	replaced by	P1365	replaces
P1376	capital of	P36	capital
P1441	present in work	P674	characters
P3373	sibling	P3373	sibling

Table 10: DocRED relations with an inverse relation. For each relation, we report its Wikidata ID, its textual description ("Name") and the Wikidata ID ("Inverse ID") and description ("Inverse Name") of its inverse relation. Note that inverse relations also include symmetric relations.

Wikidata ID	Name	Max Cardinality
P19	place of birth	2
P20	place of death	2
P22	father	2
P25	mother	2
P569	date of birth	2
P570	date of death	2
P571	inception	2
P576	dissolved, abolished or demolished	2
P580	start time	2
P582	end time	2
P585	point in time	2

Table 11: DocRED relations with bounded maximum cardinality. For each relation, we report its Wikidata ID, its textual description ("Name") and the Wikidata ID ("Inverse ID") and description ("Inverse Name") of its inverse relation. Note that inverse relations also include symmetric relations (Wikidata ID = Inverse ID).