Boosting Document-Level Relation Extraction by Mining and Injecting Logical Rules

Shengda Fan[1]*, **Shasha Mo**[1]†, **Jianwei Niu**[2,3]
[1] School of Cyber Science and Technology, Beihang University, Beijing 100191, China
[2] Zhongguancun Laboratory, Beijing 100191, China
[3] State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, Beijing 100191, China
{fanshengda, moshasha, niujianwei}@buaa.edu.cn

Abstract

Document-level relation extraction (DocRE) aims at extracting relations of all entity pairs in a document. A key challenge to DocRE lies in the complex interdependency between the relations of entity pairs. Unlike most prior efforts focusing on implicitly powerful representations, the recently proposed LogiRE (Ru et al., 2021) explicitly captures the interdependency by learning logical rules. However, LogiRE requires extra parameterized modules to reason merely after training backbones, and this disjointed optimization of backbones and extra modules may lead to sub-optimal results. In this paper, we propose MILR, a logic enhanced framework that boosts DocRE by **M**ining and **I**njecting **L**ogical **R**ules. MILR first mines logical rules from annotations based on frequencies. Then in training, consistency regularization is leveraged as an auxiliary loss to penalize instances that violate mined rules. Finally, MILR infers from a global perspective based on integer programming. Compared with LogiRE, MILR does not introduce extra parameters and injects logical rules during both training and inference. Extensive experiments on two benchmarks demonstrate that MILR not only improves the relation extraction performance (1.1%-3.8% F1) but also makes predictions more logically consistent (over 4.5% Logic). More importantly, MILR also consistently outperforms LogiRE on both counts. Code is available at https://github.com/XingYing-stack/MILR.

1 Introduction

Document-level relation extraction (DocRE) aims to identify relations of all entity pairs in a document, playing an essential role in knowledge graph construction (Luan et al., 2018), question answering (Sorokin and Gurevych, 2017), etc. A key challenge to DocRE lies in the fact that relations of entity pairs are not isolated. Rather, there exists complex interdependency between them. Consider the example in Fig. 1, where the text just explicitly shows that *Alisher* is *Chusovitina*'s `child` and that *Bakhodir* and *Chusovitina* are `married`. But according to the general interdependency between relations, which can be formulated as logical rules listed in Fig. 1, these two facts imply numerous potential facts (e.g., *Alisher* is *Bakhodir*'s `child`).

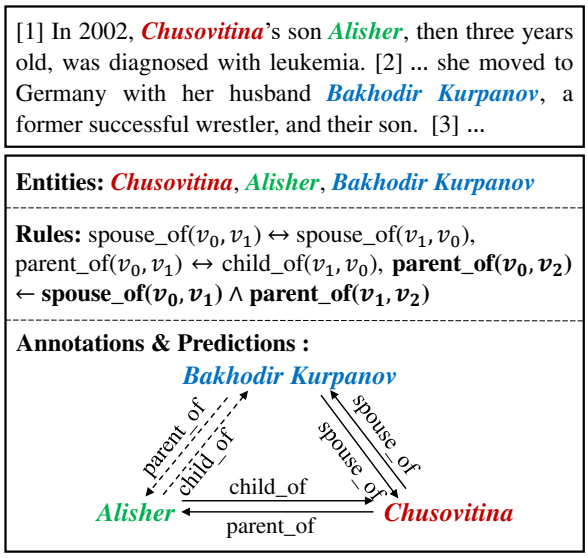

Figure 1: An example of the relational complex in DocRE and corresponding predictions produced by ATLOP (Zhou et al., 2021a). Both solid and dotted arrows represent gold annotations of relations. Nevertheless, dotted arrows represent relation facts that ATLOP cannot identify. These missing facts can be directly obtained by the bold rule together with found facts.

To capture the interdependency between entity pairs, most prior efforts focus on utilizing delicate neural networks such as pre-trained language models (Wang et al., 2019; Xu et al., 2021) or graph neural networks (Peng et al., 2017; Sahu et al., 2019; Zeng et al., 2020) to learn powerful representations. Despite their great success, these models are less transparent and still prone to making mistakes

*The first two authors contributed equally.
†Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10311 - 10323
December 7-11, 2022

when logical reasoning is needed. For example, Fig. 1 also visualizes predictions from a state-of-the-art DocRE model, ATLOP (Zhou et al., 2021a). We can see that ATLOP only extracts apparent facts such as $\texttt{spouse\_of}(Chusovitina, Bakhodir)$ while failing to identify potential facts such as $\texttt{parent\_of}(Bakhodir, Alisher)$. In fact, such potential facts can be easier identified by explicitly considering logical rules between relations (e.g., $\texttt{parent\_of}(v_0, v_2) \leftarrow \texttt{spouse\_of}(v_0, v_1) \wedge \texttt{parent\_of}(v_1, v_2)$). With this in mind, LogiRE (Ru et al., 2021) proposes to generate logical rules based on output logits of trained DocRE models (i.e., backbones) and re-extract relations by reasoning over rules [1]. However, LogiRE requires extra parameterized modules to reason merely after training backbones, and this disjointed optimization of backbones and extra modules may lead to sub-optimal results. For example, LogiRE cannot endow backbones with the sense of logical consistency during training and may cause error accumulation (See more details in Sec. 3.4).

To this end, we propose a general framework MILR to boost DocRE by **M**ining and **I**njecting **L**ogical **R**ules. Due to the lack of well-marked logical rules, MILR first mines logical rules based on conditional relative frequencies evaluated on the training set. Then consistency regularization is leveraged as an auxiliary loss to penalize the training instances that violate mined rules. Consistency regularization and commonly used classification loss are combined together to train backbones. Finally, MILR adopts a global inference method based on 0-1 integer programming, which can be seen as an extension to the widely used threshold-based inference method under logical constraints. In this manner, without training extra modules, MILR enables backbones to consider the training and predictions of all relations as a whole, explicitly capturing the interdependency between relations and thus enjoying better interpretation. Our main contributions are listed as follows:

- We propose a data-driven method to directly mine logical rules from relational annotations without needing extra resources or parameters.
- We propose a regularization loss and a programming-based inference method to constrain the output of backbones by logical rules during training and inference.
- Extensive experiments on two benchmark datasets show that MILR achieves consistent improvements on various backbones and also outperforms LogiRE in terms of relation extraction performance and logical consistency.

[1] Note that LogiRE allows in principle joint optimization of backbones and extra modules. But only the disjointed version was implemented by Ru et al. (2021) for efficiency and analyzed in this paper.

2 Preliminaries

In this section, we first present the formulation of DocRE and define two concepts: atoms and logical rules. Since MILR can be combined with various backbones, we summarize the paradigm of them.

Problem Formulation Given a document d containing n named entities $\{e_i\}_{i=1}^n$, the task of DocRE is to predict the relation types between entity pairs $(e_h, e_t)_{h,t \in \{1,\cdots,n\}, h \neq t}$. The set of relation types is defined as $\mathcal{R} \cup \{\text{NA}\}$, where $\mathcal{R}$ is a pre-defined set and NA stands for "no relation".

Atoms and Rules An *atom* (e_h, r, e_t) (or $r(e_h, e_t)$) is a binary variable indicating whether the relation r holds between the head entity e_h and the tail entity e_t. If r exists, $(e_h, r, e_t) = 1$. Otherwise $(e_h, r, e_t) = 0$. A *rule* is a formula having the conjunctive form:

$$r_{head}(v_0, v_\ell) \leftarrow r_0(v_0, v_1) \wedge \cdots \wedge r_{\ell-1}(v_{\ell-1}, v_\ell), \tag{1}$$

where $r_{head}, r_0, \cdots, r_\ell \in \widetilde{\mathcal{R}}$, $\widetilde{\mathcal{R}} = \mathcal{R} \cup \{r^{-1} | r \in \mathcal{R}\}$, $v_0, \cdots, v_\ell$ are entity variables indicating any entity, ℓ is the length of this rule. $r_{head}(v_0, v_\ell)$ and $r_{i-1}(v_{i-1}, v_i)_{i \in \{1 \cdots, \ell\}}$ are named after the head atom and body atoms, respectively. We apply the setting of probabilistic soft logic (Kimmig et al., 2012; Bach et al., 2017), assigning each rule an attribute $confidence$ in $[0, 1]$ interval. A rule R can be thought of as a prototype that can be instantiated (and denoted as $\phi(R)$) by mapping $v_0, \cdots, v_\ell$ from variables to specific entities $e_0, \cdots, e_\ell$. If all body atoms of $\phi(R)$ hold, we call $\phi(R)$ a *prediction* drawn from R, i.e., predicting the head atom holds due to R. Note that an absurd rule may have no corresponding *prediction* because its body atoms cannot hold simultaneously.

Paradigm of Backbones A typical DocRE model $\mathcal{F}$ calculates logits $\mathcal{F}_r(e_h, e_t)$ of atoms $(e_h, r, e_t)_{h,t \in \{1,\cdots,n\}, h \neq t, r \in \mathcal{R}}$. $\mathcal{F}_r(e_h, e_t)$ applied with the sigmoid function predicts whether the relation r holds for (e_h, e_t), given by

$$P(r|e_h, e_t) = \sigma(\mathcal{F}_r(e_h, e_t)), \tag{2}$$

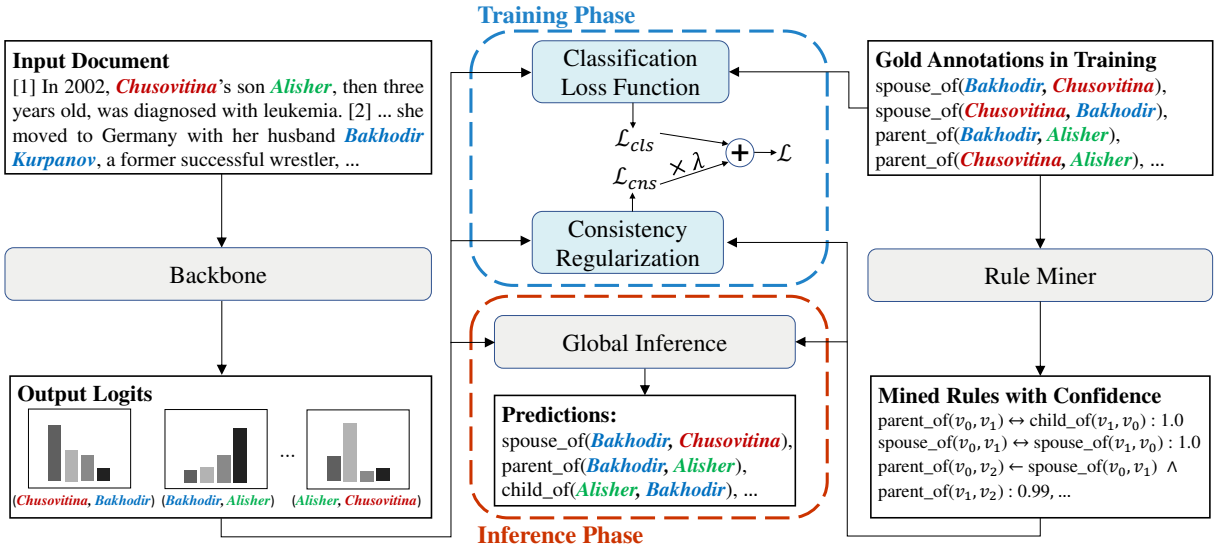

Figure 2: The overall architecture of MILR. The left part illustrates that logits are obtained by a specific backbone. The two right parts show that MILR first mines rules and then injects them into the training and inference.

where $\sigma(\cdot)$ is the sigmoid function.

To train the model, loss functions for classification (e.g., binary cross-entropy (BCE) loss or adaptive thresholding loss (Zhou et al., 2021a)) are utilized to calculate the objective (i.e., $\mathcal{L}_{cls}$).

During inference, $\mathcal{F}$ obtains the predicted relations of (e_h, e_t) by thresholding the predicted probabilities:

$$I_r(e_h, e_t) = \mathbb{I}(P(r|(e_h, e_t)) > TH_r(e_h, e_t)), \tag{3}$$

where $I_r(e_h, e_t) = 1$ means that (e_h, r, e_t) exists in predicted facts, and vice versa, $\mathbb{I}(\cdot)$ refers to the indicator function, and $TH_r(e_h, e_t)$ is the classification threshold for (e_h, r, e_t). Common threshold-based inference methods include global thresholding (Yao et al., 2019; Zeng et al., 2020) and adaptive thresholding methods (Zhou et al., 2021a; Yang Zhou, 2022). The key difference between the above two methods is whether $TH_r(e_h, e_t)$ is independent of (e_h, r, e_t).

3 Methodology

In this paper, we propose MILR as a model-agnostic framework to endow existing DocRE models with the sense of logical consistency during training and inference. The core idea of MILR is: *Both the output logits and final predictions should be constrained by logical rules.* However, most datasets do not include gold logical rules. So MILR adopts a data-driven approach to mine rules directly from relational annotations (Sec. 3.1). During training, the consistency regularization encourages the backbone to output logits that conform to mined rules (Sec. 3.2). During inference, logits along with mined rules are combined to make global predictions (Sec. 3.3). Finally, Sec. 3.4 presents a detailed comparison between MILR and LogiRE.

3.1 Rule Mining

Inspired by related work on knowledge bases and knowledge graphs (Agrawal et al., 1993; Galárraga et al., 2013), MILR takes a simple but effective frequency-based approach to mine logical rules. Intuitively, if a rule does reflect the dependencies between relations, such as $\texttt{child\_of}(v_0, v_1) \leftarrow \texttt{parent\_of}(v_1, v_0)$, its instantiated head atoms tend to co-occur with corresponding body atoms. Moreover, the *confidence* of a rule can be estimated by the conditional probability that the head atom holds when the body atoms hold.

Formally, this paper adopts the Closed World Assumption (CWA) (Reiter, 1977), any atom not in relational annotations is deemed a counterexample. Under CWA, if the head atom of a prediction $\phi(R)$ is in annotations, we call $\phi(R)$ a *true prediction*. Otherwise, it is called a *false prediction*. The *confidence* of a rule R is defined as the proportion of *true predictions* out of all *predictions*:

$$conf = \frac{support(R)}{support(R) + counter(R)}, \tag{4}$$

where $conf$ is the abbreviation of $confidence(R)$, the same below, $support(R)$ and $counter(R)$ are the number of *true predictions* and *false predictions* of rule R in the training set, respectively. Eq. 4

Algorithm 1 Rule Minier

Input: training set's annotations: Ann_{train}, expanded relation set: $\widetilde{\mathcal{R}}$, maximum rule length: $maxL$, minimum confidence: $minC$
Output: Mined rules with confidence : $rules$

1: $rules \leftarrow \{\}$
2: **for** ℓ in $1, \cdots, maxL$ **do**
3: **for** $r_{head}, r_1, \cdots, r_\ell$ in $\widetilde{\mathcal{R}}$ **do**
4: $R \leftarrow (r_{head}(v_0, v_\ell) \leftarrow r_0(v_0, v_1) \wedge \cdots \wedge r_{\ell-1}(v_{\ell-1}, v_\ell))$
5: Compute $confidence(R)$ with Ann_{train}
6: **if** $minC \leq confidence(R)$ **then**
7: $rules$.add($R : confidence(R)$)
8: **end if**
9: **end for**
10: **end for**
11: **return** $rules$

can be seen as calculating the conditional relative frequency to estimate the conditional probability. Note that if a rule R has no $prediction$, $conf$ is set to 0.

The Rule Miner (RM) takes as input the training set's annotations Ann_{train}, expanded relation set $\widetilde{\mathcal{R}}$, maximum rule length $maxL$ for constructing rules, and minimum confidence $minC$ for filtering absurd rules. As shown in Algorithm 1, RM enumerates all possible rules (Line 2-4). During enumerating, RM calculates $conf$ as in Eq. 4 (Line 5). If $conf$ is higher than $minC$, RM adds R and corresponding $conf$ to output (Line 6-7).

3.2 Consistency Regularization

After getting logical rules with confidence, a key technical challenge is how to unify discrete constraints with existing DocRE models' loss-driven learning paradigm. Inspired by the product t-norm (Gupta and Qi, 1991), we first define a rule R's ideal probability form as

$$P(r_{head}|(v_0, v_\ell)) \geq conf \cdot \eta_\ell \prod_{i=0}^{\ell-1} P(r_i|(v_i, v_{i+1})), \tag{5}$$

where ℓ is the length of R, $\eta_\ell \in [0, 1]$ is a hyper-parameter related to the rule's length ℓ for slack, and $P(r_i(v_i, v_{i+1}))$ is the output probability given by Eq. 2 [2]. Intuitively, if a rule has high confidence (near to 1), $P(r_{head}(e_0, e_\ell))$ should be greater than or at least equal to body atoms' joint probability, which is modeled as $\prod_{i=0}^{\ell-1} P(r_i(v_i, v_{i+1}))$ for simplicity. The intuition is that the rule's head atom can be deduced by corresponding body atoms or other ways, such as clear context or other rules sharing the same head atom. With $conf$ dropping, this constraint becomes more relaxed.

However, without regularization, above ideal probabilistic forms of rules are likely to be broken during the training of backbones, especially when head atoms' relational types are uncommon (Huang et al., 2022). So this paper argues besides DocRE models' vanilla classification loss $\mathcal{L}_{cls}$, there is another loss $\mathcal{L}_{cns}$ related to logical consistency that should be minimized. To put both $\mathcal{L}_{cls}$ and $\mathcal{L}_{cns}$ in the log space of probabilities, given a document d, we formulate $\mathcal{L}_{cns}$ as

$$L_{cns} = \sum_{R \in rules} \sum_{e_i \in d} \max(0, \log(\eta_\ell \cdot conf) + \sum_{i=0}^{\ell-1} \log(P(r_i|(e_i, e_{i+1}))) - \log(P(r_{head}|(e_0, e_\ell)))). \tag{6}$$

L_{cns} enumerates all instantiated rules and regularize corresponding logits to satisfy the ideal forms defined in Formula 5. If rules' ideal probabilistic forms are nearly satisfied, the consistency regularization loss L_{cns} and its gradient are both small, so they have little impact on backbones' training. If not, L_{cns} would incur a large magnitude of gradients in training, which regularize backbones to satisfy logical consistency.

To sum up, the training objective in MILR is

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \cdot \mathcal{L}_{cns}, \tag{7}$$

where λ is a hyper-parameter to balance two losses. In this way, the learning process seeks to unify the likelihood nature of individual atoms and the logical nature between multiple relations, therefore supporting the backbone to comprehensively understand given annotations.

3.3 Global Inference

Although logical rules have been injected during training, backbones may still output predictions that violate logical rules during inference. Motivated by this observation, MILR leverages a programming-based method to inject logical rules during inference, forming a global inference method. Note that this method can be seen as an extension to threshold-based methods mentioned in

[2] For any reverse relation type $r^{-1} \in \widetilde{\mathcal{R}} \setminus \mathcal{R}$, we define $P(r^{-1}|(v_i, v_{i+1})) = P(r|(v_{i+1}, v_i))$.

Eq. 3. To make it clearer, we first revisit threshold-based methods from the perspective of 0-1 integer programming:

Fact 1. *Let $\mathcal{F}$ be a DocRE model, $\mathcal{F}_r(e_h, e_t)$ be the output logits, $TH_r(e_h, e_t)$ be the thresholds, $I_r(e_h, e_t)$ be the predicted results of atoms $(e_h, r, e_t)_{h,t\in\{1,\cdots,n\},h\neq t,r\in\mathcal{R}}$, $g_r(e_h, e_t) = \sigma(\mathcal{F}_r(e_h, e_t) - \sigma^{-1}(TH_r(e_h, e_t)))$. An optimal solution for the following problem:*

$$\begin{aligned}\min_{I_r(e_h,e_t)} & -\sum_{h\neq t}\sum_{r\in\mathcal{R}}\big(I_r(e_h,e_t)\log g_r(e_h,e_t) \\ & + (1-I_r(e_h,e_t))\log(1-g_r(e_h,e_t))\big) \\ s.t. \quad & I_r(e_h,e_t)\in\{0,1\},\end{aligned} \tag{8}$$

is $I^*_r(e_h, e_t) = \mathbb{I}(P(r|(e_h, e_t)) > TH_r(e_h, e_t))$ where $P(r|e_h, e_t) = \sigma(\mathcal{F}_r(e_h, e_t))$.

The proof is in Appendix A. The construction of the objective function is inspired by the BCE loss function. Thus, threshold-based methods can be seen as leveraging potential prediction results $I_r(e_h, e_t)$ as binary decision variables to unconstrainedly minimize the sum of cross-entropy of the distribution $(g_r(e_h, e_t), 1 - g_r(e_h, e_t))$ relative to the distribution $(I_r(e_h, e_t), 1 - I_r(e_h, e_t))$.

This perspective inspires us to naturally inject logical rules as the programming problem's constraints. Intuitively, for a rule R, logical consistency requires that its predicted body atoms all hold, then its predicted head atom also holds. If any body atom fails, logical consistency has no constraints on the predicted head atom. This can be mathematically expressed as $\forall e_i,\ \sum_{i=0}^{\ell-1} I_{r_i}(e_i, e_{i+1}) \leq I_{r_{head}}(e_0, e_l) + \ell - 1$. Adding these logical constraints and symmetry constraints could get the vanilla form of the global inference method:

$$\begin{aligned}\min_{I_r(e_h,e_t)} & -\sum_{h\neq t}\sum_{r\in\mathcal{R}}\big(I_r(e_h,e_t)\log g_r(e_h,e_t) \\ & + (1-I_r(e_h,e_t))\log(1-g_r(e_h,e_t))\big) \\ s.t. \quad & I_r(e_h,e_t)\in\{0,1\} \\ & \forall r^{-1}\in\widetilde{\mathcal{R}}\setminus\mathcal{R}, I_{r^{-1}}(e_h,e_t) = I_r(e_t,e_h) \\ & \forall R\in rules, \forall e_i,\ \sum_{i=0}^{\ell-1} I_{r_i}(e_i,e_{i+1}) - \ell + 1 \\ & \qquad\qquad \leq I_{r_{head}}(e_0,e_\ell).\end{aligned} \tag{9}$$

This vanilla form can be seen as leveraging inference results to minimize the BCE loss under logical constraints, which is also the idea of the training objective defined in Formula 7. Note that this vanilla form can be solved through the branch and bound method (Lawler and Wood, 1966) or an off-the-shelf optimizer such as Gurobi (Gurobi Optimization, LLC, 2022).

However, this problem has $O(n^{maxL+1}\cdot|rules|)$ logical constraints where n is the number of entities. Redundant constraints make the calculation terribly slow. To alleviate this problem, we propose a heuristic strategy to simplify constraints as in Algorithm 2. As seen, we only add logical constraints on *predictions* whose body atoms are all predicted to be true by the threshold-based approach. Intuitively, this strategy can be viewed as revising some body atoms and corresponding head atoms with logical rules, while the predicted results of other atoms are the same as those in the silver labels produced by thresholding probabilities. Mathematically, this strategy can also be viewed as an approximation of the positive constraints at the optimal solution (Forsgren et al., 2016). In this way, the number of constraints can be significantly reduced since most entity pairs show "no relation".

Algorithm 2 Simplifying Logical Constraints

Input: backbone's output logits: $\mathcal{F}_r(e_h, e_t)$, classification thresholds: $TH_r(e_h, e_t)$, mined rules: $rules$

Output: simplified logical constraints: SLC

```
Get silver labels by thresholding probabilities as in Eq. 3.
SLC ← {}
for R in rules do
    Get R's all predictions Pr(R) using silver labels as ground truth.
    for φ(R) in Pr(R) do
        SLC.add(Σ_{i=0}^{ℓ-1} I_{r_i}(e_i, e_{i+1}) − ℓ + 1 ≤ I_{r_head}(e_0, e_ℓ))
    end for
end for
return SLC
```

When evaluating the models, we find that adding compensation terms to form the objective function can further improve the performance. The modified objective function is given by

$$\begin{aligned}\min_{I_r(e_h,e_t)} & -\sum_{h\neq t}\sum_{r\in\mathcal{R}}\big(I_r(h,t)\cdot\log g_r(e_h,e_t) + \\ & (1-I_r(h,t))\cdot\tau\cdot(-\log p_r)^k\cdot\log(1-g_r(e_h,e_t))\big),\end{aligned} \tag{10}$$

where τ, k are hyper-parameters, p_r is the frequency of relation r evaluated on the training set.

These compensation terms can help to alleviate the class imbalance problem in DocRE.

To sum up, the final form of global inference takes Formula 10 as the objective and utilizes Algorithm 2 to construct the set of logical constraints. Based on integer programming, the logical inconsistencies with low probabilities can be filtered out, leading to better performance and interpretability.

3.4 Comparison with LogiRE

Although LogiRE (Ru et al., 2021) and MILR share the same purpose of injecting logical rules into backbones, MILR has three advantages. Firstly, MILR is more efficient without training extra modules. Secondly, MILR leverages consistency regularization to endow backbones with the sense of logical consistency during training. In contrast, LogiRE does not touch the training process. Therefore, backbones under LogiRE are more sensitive to noisy labels, which is relatively common in DocRE (Huang et al., 2022). Thirdly, MILR can address more error types, categorized by where the error occurs in logical rules. Through a programming-based approach during inference, MILR can theoretically alleviate False Negatives of Head atoms (FNH) and False Positives of Body atoms (FPB) [3]. In contrast, LogiRE can only handle FNH because LogiRE computes the final logits of atoms to be evaluated through meta-paths, characterized by the backbones' misleading logits. When LogiRE faces FPB (i.e., backbones output high logits for triples that do not hold), LogiRE will uncritically treat these logits as true positives and introduce more False Positives of Head atoms (FPH). Note that both MILR and LogiRE cannot deal with FPH and False Negatives of Body atoms (FNB) because there is nothing to reason about, and the logical constraint has been satisfied. For clarity, we summarize the above discussions in Table 1.

Frameworks	FPH	FNH	FPB	FNB
MILR	✗	✓	✓	✗
LogiRE	✗	✓	✗	✗

Table 1: MILR vs. LogiRE on processable error types.

[3] In the experiment, MILR may inevitably transform true positives to false negatives or true negatives to false positives.

4 Experiments

4.1 Experimental Setups

Datasets. The experiments are conducted on DWIE (Zaporojets et al., 2021) and DocRED (Yao et al., 2019). DWIE is a human-annotated dataset for document-level information extraction including DocRE. Besides relational annotations, DWIE also provides hand-crafted logical rules for constructing the dataset. DocRED is a large-scale and widely used dataset for DocRE. However, a recent study by Huang et al. (2022) finds many false negative samples in the original DocRED. For a fairer comparison, we utilize their public relabeled dataset as the test set. The details of the above two datasets are listed in Appendix B.

Metrics. Following Yao et al. (2019) and Ru et al. (2021), we utilize F1, Ign F1, and Logic as metrics, where F1 & Ign F1 are for relation extraction and Logic is for logical consistency. The calculation of Ign F1 excludes relational facts appearing both in the training set and test set, avoiding leakages of the test set. Logic is utilized to evaluate whether the predictions satisfy gold rules. Note that pre-defined rules used in calculating Logic are not included in the training and inference of MILR.

Baselines. We utilize the following models as baselines and backbones: LSTM & BiLSTM (Yao et al., 2019), GAIN (Zeng et al., 2020), and ATLOP (Zhou et al., 2021a). Among these backbones, ATLOP is re-evaluated as state-of-the-art by Huang et al. (2022) and serves as the main backbone in latter discussion. These baselines have various encoders, such as sequence- and graph-based neural networks, pre-trained language models, and attention mechanisms. These backbones also include various loss functions such as BCE and adaptive-thresholding loss. We are convinced that such a setting can lead to a more comprehensive evaluation. We also compare MILR with LogiRE under different backbone settings.

Experimental Settings. We utilize the public repositories of backbones to implement our experiments[4,5,6,7]. For a fair comparison, we rerun backbones following the recommended hyperparameters and report their performance with and without LogiRE & MILR. We report the median results of five runs using different random seeds

[4] https://github.com/thunlp/DocRED
[5] https://github.com/DreamInvoker/GAIN
[6] https://github.com/wzhouad/ATLOP
[7] https://github.com/rudongyu/LogiRE

Model	Dev			Test		
	Ign F1	F1	Logic	Ign F1	F1	Logic
LSTM	31.71	38.35	64.15	31.65	41.42	62.27
LSTM + LogiRE	32.02(+0.31)	38.48(+0.13)	77.93(+13.78)	32.58(+0.93)	42.03(+0.61)	73.01+10.74)
LSTM + MILR	33.12(+1.41)	39.95(+1.60)	78.84(+14.69)	33.75(+2.10)	43.35(+1.93)	74.39(+12.12)
BiLSTM	32.14	39.66	52.24	33.88	43.54	60.53
BiLSTM + LogiRE	32.39(+0.25)	40.32(+0.66)	69.24(+17.00)	34.21(+0.33)	43.95(+0.45)	73.13(+12.60)
BiLSTM + MILR	34.05(+1.91)	41.22(+1.56)	74.62(+22.38)	35.09(+1.21)	44.65(+1.11)	73.92(+13.39)
GAIN	58.89	63.81	85.25	61.36	67.45	86.85
GAIN + LogiRE	58.98(+0.09)	64.90(+1.09)	91.25(+6.00)	61.58(+0.22)	68.71(+1.26)	91.71+4.86)
GAIN + MILR	61.22 (+2.33)	65.85(+2.04)	93.77(+8.52)	62.77 (+1.41)	69.23(+1.78)	91.92(+5.07)
ATLOP	63.37	69.87	86.14	67.29	75.13	88.62
ATLOP + LogiRE	64.54(+1.17)	70.66(+0.79)	90.33(+4.19)	68.13(+0.84)	75.67(+0.54)	91.42(+2.80)
ATLOP + MILR	**67.18**(+3.81)	**72.05**(+2.97)	**94.85**(+8.71)	**69.84**(+2.55)	**76.51**(+1.38)	**93.16**(+4.54)

Table 2: Main results on DWIE (%). **Bold** indicates the best performance.

for all experiments. For GAIN and ATLOP, BERT-base-uncased (Devlin et al., 2019) is used as the pre-trained language model [8]. With regard to hyper-parameters introduced by MILR, we tune λ from {1e-3, 3e-3, 5e-3, 1e-2}, tune τ from {0.5, 0.8, 1.0, 2.0} and tune k from {0, 0.5, 1.0, 1.5, 1.8, 2.0}. All hyper-parameters are chosen based on the F1 score on the development set. All models are implemented in PyTorch (Paszke et al., 2019) and trained on one Tesla V100 GPU. We provide detailed hyper-parameter settings in Appendix D.

4.2 Results and Discussions

Results on DWIE. The experimental results on DWIE are reported in Table 2. By explicitly considering logical rules between different relations, LogiRE and MILR improve the relation extraction performance on all backbones. Moreover, our MILR framework consistently outperforms LogiRE by a large margin. The improvements demonstrate the strong generality ability and effectiveness of MILR, which is compatible with a variety of encoders and loss functions. With ATLOP, MILR achieves a state-of-the-art Ign F1 of 69.84% and F1 of 76.51%. In addition, MILR improves the Logic scores by 4.5%-22.4%, consistently outperforming LogiRE. The improved Logic scores demonstrate the effectiveness of MILR in making predictions more consistent with gold rules.

Results on DocRED. Following LogiRE, only evaluation results of strong baselines are included in Table 3. As seen, MILR consistently outperforms LogiRE by an average of 1.60% in Ign F1 and 1.64% in F1. The performance gaps are more significant than on DWIE, mainly due to the difference in the training of backbones. Consistency regularization in MILR can help backbones not to overfit noisy (i.e., false negative) labels. On the other hand, backbones in LogiRE only adopt the likelihood-based training strategy resulting in overfitting. LogiRE further utilizing logits output by overfitted backbones as gold features leads to sub-optimal results.

[8] Note that this setup differs from that of Ru et al. (2021), so the reported baseline performance may vary.

Model	Test	
	Ign F1	F1
GAIN	41.26	41.68
GAIN + LogiRE	41.53(+0.27)	41.89(+0.21)
GAIN + MILR	42.89(+1.63)	43.17(+1.49)
ATLOP	41.67	41.95
ATLOP + LogiRE	42.47(+0.80)	42.73(+0.78)
ATLOP + MILR	**44.30**(+2.63)	**44.72**(+2.77)

Table 3: Main results on DocRED (%).

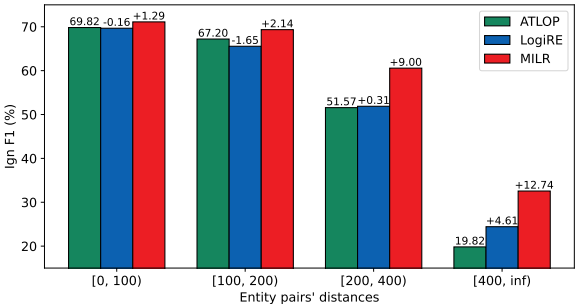

Figure 3: Ign F1 comparison between ATLOP, ATLOP + LogiRE and ATLOP+MILR with different distances.

Results of Mined Rules. Using $minC$ and $maxL$ listed in Appendix D, the proposed RM mines 90 and 168 rules for DWIE and DocRED, respectively. The mined rules include symmetry, impli-

cation, composition, and others. Also, RM can mine 40 out of 41 gold rules annotated in DWIE. The only missing one is $\texttt{event\_in0}(v_0, v_2) \leftarrow \texttt{event\_in1}(v_0, v_1) \wedge \texttt{in0}(v_1, v_2)$ because the relation `event_in1` does not exist in the training set. The large scale and high quality of mined rules prove the effectiveness of the proposed RM. The case study of mined rules is in Appendix C.

Performance with respect to entity pairs' distances. To demonstrate the power of logical rules in capturing long-range dependencies, we break down the relation extraction performance into four groups according to the distance between entity pairs. Following Ru et al. (2021), the distance of an entity pair is calculated as the number of tokens between the nearest mentions. Fig. 3 presents corresponding results of different models. As shown, all models' performance degrades with the growth of distance, indicating the difficulty of modeling long-term dependencies. However, MILR consistently outperforms the other two baselines in all four groups. Furthermore, the performance gains of combining MILR increase as distance grows. For distances falling into $[200, 400)$ and $[400, \inf)$, MILR achieves 9.00% and 12.74% Ign F1 enhancement, respectively. These results demonstrate the superiority of MILR in incorporating rules, which could go beyond noisy text and directly capture the high-level interdependency between relations.

To further investigate the performance difference between LogiRE and MILR, we plot precision and recall in Fig. 4 with different entity pairs' distances. The results show that while LogiRE slightly outperforms in recall, MILR performs much better in precision. This difference sheds light on why MILR performs better overall. In fact, this is precisely related to the different behavior patterns stated in Sec. 3.4. Reasoning over backbones' logits as gold features, LogiRE only can complement numerous poor-quality facts, resulting in high recall and low precision. In contrast, MILR adds a few higher-quality facts and filters out several FPB. A qualitative comparison example is in Appendix C.

Efficiency comparison. We benchmark the size of parameters and running time of ATLOP, LogiRE, and MILR on one Tesla V100 GPU. Unlike MILR, LogiRE introduces ∼5.8 M additional parameters compared to ATLOP. As for training time, MILR yields an extra ∼2.8 minutes per training epoch, while LogiRE yields an extra ∼104.8 minutes per iteration to train additional modules. In terms of inference time, MILR generates an extra ∼3.1 minutes per epoch, while the extra time introduced by LogiRE is negligible. Overall, MILR is more time- and space-efficient than LogiRE.

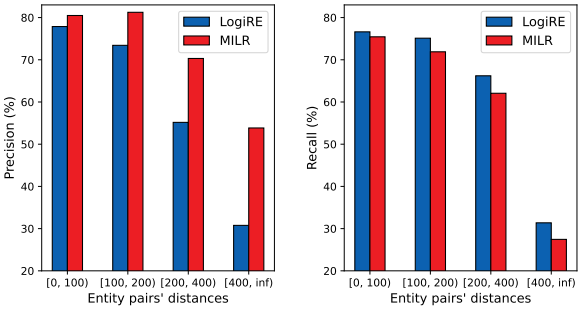

Figure 4: Precision and recall comparison between LogiRE and MILR with different distances.

Ablation study. We perform an ablation experiment to validate the effectiveness of MILR's two components. Table 4 provides the results when each component is excluded at a time, where CR and GI denote the consistency regularization and the global inference, respectively. We observe that both variants excluding one of the components still outperform the backbone, indicating that both components are beneficial. Furthermore, the results show that CR and GI are not mutually replaceable, suggesting that injecting logical rules into the training and inference is equally vital. We also analyze the effect of simplifying constraints in Appendix E.

Model	Dev		Test	
	Ign F1	F1	Ign F1	F1
ATLOP+MILR	**67.18**	**72.05**	**69.84**	**76.51**
- CR	63.82	70.27	68.87	76.20
- GI	66.74	71.39	69.29	76.05
ATLOP	63.37	69.87	67.29	75.13

Table 4: Ablation study on the DWIE dataset by using ATLOP as the backbone (%).

5 Related Work

Document-level relation extraction. Previous efforts on DocRE mainly focus on learning better representations. Powerful neural networks, such as attention mechanisms (Yao et al., 2019; Zhou et al., 2021a), graph neural networks (Christopoulou et al., 2019; Zhang et al., 2020; Zeng et al., 2020) and pre-trained language models (Wang et al., 2019; Tang et al., 2020; Xu et al., 2021) are utilized as the encoder to generate representations of entity pairs. For another, considering the severe class im-

balance problem in DocRE, especially with many samples expressing "no relation", some studies aim at designing instance-dependant thresholds and corresponding loss functions (Zhou et al., 2021a; Yang Zhou, 2022). Unlike previous work, we propose MILR as a general framework that can be combined with various encoders and loss functions. MILR proposes to constrain the output of DocRE models with logical rules, thus improving their interpretability and overall performance.

Deep learning with logical rules. In recent research on deep learning, logical rules have been applied to various topics such as model interpretability (Hu et al., 2016), knowledge base construction (Demeester et al., 2016; Ding et al., 2018), natural language inference (Li and Srikumar, 2019) and sentiment analysis (Deng and Wiebe, 2015). As for information extraction, using hand-crafted rules, Wang and Pan (2020); Zhou et al. (2021b) achieve great success on sentence-level or clinical tasks. Unlike the above two studies, our work does not require any additional information. Also, as far as we know, only LogiRE (Ru et al., 2021) attempts to incorporate logical rules in DocRE. Compared with LogiRE, MILR is lighter and more powerful.

6 Conclusion

In this paper, we propose a novel framework MILR to explicitly capture the interdependency of relations between different entity pairs in DocRE. MILR mines logical rules from relational annotations and utilizes rules to constrain the output logits and final predictions of backbones. The proposed MILR is a general framework and has shown to be effective with various base models, consistently improving their effectiveness and logical consistency.

Limitations

Although making some progress, our MILR framework still has several limitations. First of all, the rule miner in MILR adopts the closed world assumption, which heavily relies on the scale and quality of annotations. If the annotation size is small or has many wrong labels, the $confidence$ estimated by the rule miner will not be an accurate measure of the rule's reliability. And these inaccurately estimated rules would mislead the following training and inference. In addition, the global inference method in MILR requires off-the-shelf optimizers to solve programming problems, which results in additional CPU consumption and computation time. Moreover, the global inference method still relies on the quality of backbones' logits, i.e., wrongly estimated atoms may cause error propagation. Last but not least, MILR injects logical rules only once to enhance backbones during training, ignoring the interactive nature between embedding and logical reasoning, which may lead to sub-optimal results (Cheng et al., 2021; Zhao et al., 2022). We will make up for the above deficiencies in future work.

Ethics Statement

Compared with sentence-level counterparts, DocRE models, including proposed MILR, have greater potential for analyzing large volumes of online text and mining private information among different users. Aware of this concern, all data used in this paper is public and does not involve private information. Also, the proposed framework should not be used to analyze any information involving personal privacy in the future.

Acknowledgements

This work was supported by National Natural Science Foundation of China (62106013). We sincerely thank Yuerong Li, Zhengsu Chen and other anonymous reviewers for their insightful and detailed comments.

References

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216.

Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18:1–67.

Kewei Cheng, Ziqing Yang, Ming Zhang, and Yizhou Sun. 2021. UniKER: A unified framework for combining embedding and definite horn rule reasoning for knowledge graph inference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9753–9771. Association for Computational Linguistics.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. 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)*, pages 4925–4936.

Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. 2016. Lifted rule injection for relation embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1389–1399.

Lingjia Deng and Janyce Wiebe. 2015. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 179–189.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Boyang Ding, Quan Wang, Bin Wang, and Li Guo. 2018. Improving knowledge graph embedding using simple constraints. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 110–121.

Anders Forsgren, Philip E Gill, and Elizabeth Wong. 2016. Primal and dual active-set methods for convex quadratic programming. *Mathematical programming*, (1):469–508.

Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian M. Suchanek. 2013. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 413–422.

Madan M Gupta and J11043360726 Qi. 1991. Theory of t-norms and fuzzy inference methods. *Fuzzy sets and systems*, (3):431–450.

Gurobi Optimization, LLC. 2022. Gurobi Optimizer Reference Manual.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2410–2420.

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.

Angelika Kimmig, Stephen Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2012. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS workshop on probabilistic programming: foundations and applications*, pages 1–4.

Eugene L Lawler and David E Wood. 1966. Branch-and-bound methods: A survey. *Operations research*, (4):699–719.

Tao Li and Vivek Srikumar. 2019. Augmenting neural networks with first-order logic. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 292–302.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

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*, pages 101–115.

Raymond Reiter. 1977. On closed world data bases. In *Logic and Data Bases, Symposium on Logic and Data Bases, Centre d'études et de recherches de Toulouse, France, 1977*, Advances in Data Base Theory, pages 55–76.

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.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4309–4316.

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

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In

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

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. Hin: Hierarchical inference network for document-level relation extraction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 197–209. Springer.

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

Wenya Wang and Sinno Jialin Pan. 2020. Integrating deep learning with logic fusion for information extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9225–9232.

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.

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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14149–14157.

Wee Sun Lee Yang Zhou. 2022. None class ranking loss for document-level relation extraction. In *Proceedings of IJCAI*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, (4):102563.

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

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

Qingjuan Zhao, Jianwei Niu, and Xuefeng Liu. 2022. Als-mrs: Incorporating aspect-level sentiment for abstractive multi-review summarization. *Knowledge-Based Systems*, page 109942.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021a. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14612–14620.

Yichao Zhou, Yu Yan, Rujun Han, J. Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2021b. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14647–14655.

A Proof

Proof of Fact 1. Since neither the objective function nor the constraints involve the interaction of decision variables, the overall optimal solution can be computed by solving the following sub-problems:

$$\begin{aligned}\min_{I_{r'}(e_{h'},e_{t'})} & -I_{r'}(e_{h'},e_{t'})\log g_{r'}(e_{h'},e_{t'}) \\ & -(1-I_{r'}(e_{h'},e_{t'}))\log(1-g_{r'}(e_{h'},e_{t'})) \\ s.t. \quad & I_{r'}(e_{h'},e_{t'})\in\{0,1\}\end{aligned} \tag{11}$$

where $\forall r' \in \mathcal{R}, \forall h', t' \in \{1, 2, ..., n\}$, and $h \neq t$. This sub-problem's optimal solution can be obtained by comparing the objective function value at 0 or 1, which leads to

$$I^*_{r'}(e_{h'},e_{t'}) = \begin{cases} 0, & 2g_{r'}(e_{h'},e_{t'}) \leq 1, \\ 1, & 2g_{r'}(e_{h'},e_{t'}) > 1. \end{cases} \tag{12}$$

Considering that $\sigma(\cdot)$ is a monotonically increasing function and $\sigma(0) = 0.5$, the Formula 12 could be simplified as:

$$I^*_{r'}(e_{h'},e_{t'}) = \begin{cases} 0, & P(r'|e_{h'},e_{t'}) \leq TH_{r'}(e_{h'},e_{t'}), \\ 1, & P(r'|e_{h'},e_{t'}) > TH_{r'}(e_{h'},e_{t'}). \end{cases} \tag{13}$$

where $P(r'|e_{h'},e_{t'}) = \sigma(\mathcal{F}_{r'}(e_{h'},e_{t'}))$. Combining all sub-problems' optimal solutions could obtain the optimal solution to the original problem: $I^*_r(e_h,e_t) = \mathbb{I}(P(r|(e_h,e_t)) > TH_r(e_h,e_t))$, for $h,t \in \{1,\cdots,n\}, h \neq t$, and $r \in \mathcal{R}$.

□

B Datasets

Our framework is evaluated on two DocRE benchmark datasets, DWIE and DocRED. The dataset preprocessing on DWIE mostly follows Ru et al. (2021), except that we remove facts that share the same head and tail entity. Recently, a number of studies have found that labeling mistakes are common in DocRED (Ru et al., 2021; Xie et al., 2022; Tan et al., 2022), especially the many false negative samples. To this end, Huang et al. (2022) relabeled 96 documents in the development (dev.) set for a fairer test. In this paper, we adopt the original training set in DocRED as our training set, the original dev set excluding relabeled documents as our dev set, and the relabeled samples as our test set. More statistical information is listed in Table 5.

Dataset		#Doc.	#Rel.	#Ent.	#Facts
DWIE	train	602	65	16494	14403
	dev	98		2785	2624
	test	99		2623	2459
DocRED	train	3053	96	59493	38180
	dev	904		17685	11109
	test	96		1893	3308

Table 5: Statistics of DWIE and DocRED.

C Case Study

Case study of Mined Logical Rules We list several mined rules on the DWIE dataset in Table 6, where h, z, t are entity variables, and the same below. For simplicity, we have transformed reversed atoms (e_t, r^{-1}, e_h) into positive ones (e_h, r, e_t). We can see that these logical rules are meaningful and interpretable. The first rule is an implication rule. The second rule is symmetric. The third and fourth rules are two-hop compositional rules. This case study shows that RM can mine practical rules for subsequent injection.

Rules	Confidence
$\texttt{parent\_of}(h,t) \leftarrow \texttt{child\_of}(t,h)$	1.0
$\texttt{vs}(h,t) \leftrightarrow \texttt{vs}(t,h)$	1.0
$\texttt{citizen\_of}(h,t) \leftarrow \texttt{agent\_of}(h,z) \wedge \texttt{in0}(z,t)$	1.0
$\texttt{plays\_in}(h,t) \leftarrow \texttt{played\_by}(z,h) \wedge \texttt{character\_in}(z,t)$	1.0

Table 6: Case study of mined rules.

Case study of DocRE Fig. 5 shows several relation extraction cases of ATLOP, LogiRE and MILR, where ATLOP is used as the backbone for the last two frameworks. And there are two logical rules related to the this case study: $\texttt{citizen\_of}(h,t) \leftarrow \texttt{citizen\_of-x}(h,z) \wedge \texttt{gpe0}(z,t)$ and $\texttt{based\_in0-x}(h,t) \leftarrow \texttt{based\_in0}(h,z) \wedge \texttt{gpe0}(t,z)$. As shown, ATLOP extracts two true facts and one false fact. Moreover, these three facts can be seen precisely as instantiated body atoms of the above two rules. Thus, as discussed in Sec. 3.4, these predictions can be further complemented (through LogiRE and MILR) or filtered (through MILR) by logical rules.

The results show that LogiRE complements two head atoms, which handles a false negative and introduces one more false positive. This implies the disadvantage of disjointed training of backbones and extra modules, i.e., uncritically treating logits

[1] All eyes will be on a *Finn* and a *Spaniard* at the *German Grand Prix* at *Hockenheim* on *Sunday* as the title battle between *Kimi Raikkonen* and *Fernando Alonso* looks to put local hero ***Schumi*** in the shade.
[2] The battle between *McLaren*'s flying *Finn* *Kimi Raikkonen* and championship leader *Renault*'s speedy *Spaniard* *Fernando Alonso* is diverting eyes from the world record holding ***German*** on his own turf.
[3] "I am really looking forward to racing again this weekend in ***Germany***, "said *Montoya*, who won at *Hockenheim* for ***Williams*** two years ago and got his first victory for *McLaren* in *Britain* two weeks ago."

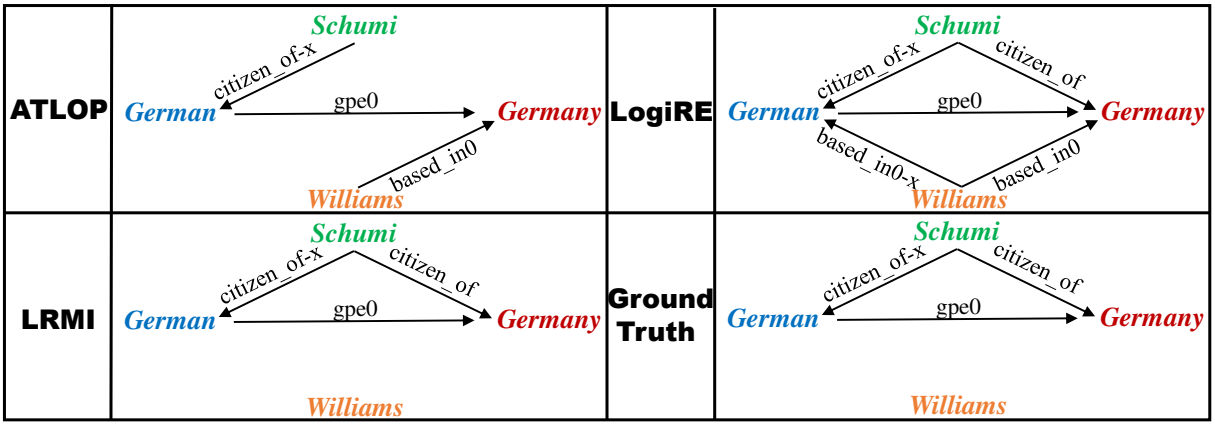

Figure 5: The case study of proposed MILR framework and baselines. For clarity, named entities involved in this case study are marked in color and other entities are underlined. The specific definition of the relations can be found in the original DWIE paper (Zaporojets et al., 2021).

Hyper-param	**DWIE**				**DocRED**	
	LSTM	BiLSTM	GAIN	ATLOP	GAIN	ATLOP
$minC$ for rule miner	0.98	0.98	0.98	0.98	0.7	0.7
$maxL$ for rule miner	2	2	2	2	2	2
λ for consistency regularization	1e-3	1e-3	3e-3	1e-3	1e-2	1e-3
η_1 for consistency regularization	$e^{-0.05}$	$e^{-0.05}$	$e^{-0.05}$	$e^{-0.05}$	$e^{-0.05}$	$e^{-0.05}$
η_2 for consistency regularization	$e^{-0.1}$	$e^{-0.1}$	$e^{-0.1}$	$e^{-0.1}$	$e^{-0.1}$	$e^{-0.1}$
τ for global inference	1.0	1.0	0.8	0.8	2.0	1.5
k for global inference	1.5	1.5	1.0	0.5	1.5	1.5

Table 7: Hyper-parameter settings on different datasets.

as gold features easily causes error propagation. Compared with LogiRE, MILR reduces the risk of error propagation through regularization during training and revising during inference. As shown, our MILR framework filters the wrongly estimated body atom and complements the missing atom.

D Hyper-Parameter Settings

Table 7 provides the detailed hyper-parameter settings in regard to different backbones and datasets.

E Effect of Simplifying Constraints

As stated in Sec. 3.3, we propose a heuristic strategy to simplify logical constraints in the global inference method. The comparison in F1, Ign F1, the average number of constraints (Con.) across all documents, and the average running time (RT) across all documents are shown in Table 8. As seen, our simplifying method could significantly reduce the number of constraints and computation time. Moreover, the drops of F1 & Ign F1 are relatively mild, showing that our simplifying method is an effective way to speed up calculating without hurting performance much.

Strategies	**#Con.**	**#RT**	**F1**	**Ign F1**
Simplified	**28.4**	**1.80** s/it	72.05	67.18
Vanilla	4.1e+06	1074 s/it	**72.26**	**67.35**

Table 8: Comparison between the vanilla and simplified form of the global inference method with ATLOP on the dev set of DWIE. All settings are the same except for the strategy of formulating logical constraints.