DRK: Discriminative Rule-based Knowledge for Relieving Prediction Confusions in Few-shot Relation Extraction

Mengru Wang, Jianming Zheng*, Fei Cai, Taihua Shao, Honghui Chen
Science and Technology on Information Systems Engineering Laboratory,
National University of Defense Technology, China
{wangmengru20,zhengjianming12,caifei08,shaotaihua13,chenhonghui}@nudt.edu.cn

Abstract

Few-shot relation extraction aims to identify the relation type between entities in a given text in the low-resource scenario. Albeit much progress, existing meta-learning methods still fall into **prediction confusions** owing to the limited inference ability over shallow text features. To relieve these confusions, this paper proposes a *discriminative rule-based knowledge* (DRK) method. Specifically, DRK adopts a logic-aware inference module to ease the word-overlap confusion, which introduces a logic rule to constrain the inference process, thereby avoiding the adverse effect of shallow text features. Also, DRK employs a discrimination finding module to alleviate the entity-type confusion, which explores distinguishable text features via a hierarchical contrastive learning. We conduct extensive experiments on four types of meta tasks and the results show promising improvements from DRK (6.0% accuracy gains on average). Besides, error analyses reveal the word-overlap and entity-type errors are the main courses of mispredictions in few-shot relation extraction.

1 Introduction

With the emergence of new relation types and ever-increasing annotation costs, traditional data-driven relation extraction (RE) methods cannot survive in this low-data regime (Wang et al., 2021; Chia et al.). Therefore, the task of few-shot relation extraction is proposed to cope with such a low-resource dilemma. In few-shot RE, meta-learning (ML) has been extensively employed and attained promising performance, the core of which is to learn the generalization ability from the data-rich classes to help predict the data-scarce classes (Han et al., 2018). These ML approaches can be roughly divided into two categories: basic ML only accessible to raw sentence text, e.g., prototype (Snell et al., 2017) and MAML (Finn et al., 2017), and knowledge-based ML with the additional external knowledge (Zheng et al., 2020a), e.g., TD-proto (Yang et al., 2020) and REGRAB (Qu et al., 2020).

Albeit much progress, the insufficient labeled data forces existing ML methods to make shallow inferences based on superficial text features, e.g., the word overlap (Utama et al., 2021) and the matched entity type (Brody et al., 2021). In this light, as shown in Fig.1, when the support instances of some relations exhibit massive word overlaps or have a matched entity-type pair, these ML methods inevitably get caught in the **prediction confusions** (Wang et al., 2020). Based on the taxonomy of similar pattern, prediction confusions can be further classified into the word-overlap confusion and the entity-type confusion as shown in Fig.1.

For the word-overlap confusion, intuitively, incorporating external knowledge, e.g., relation descriptions (Yang et al., 2020), an entity-level knowledge graph (KG) (Roy and Pan, 2021) or entity types (Sainz et al., 2021; Hao et al., 2019; Yang et al., 2021), into RE models can ease this semantic uncertainty to some extent. However, this simple incorporation (concatenation in the majority) just provides additional information, and cannot alter the fact that existing ML methods still process the shallow inferences. Even worse, the lengthened text may exceed the input limit and introduce noise, thereby degrading the model performance. Besides, the entity-type confusion can not be relieved with the introduction of external knowledge. Since external information generally contains entity type information, the simple introduction cannot work in relations whose entity-type pairs are matched in head and tail positions. Although the entity masking (Li et al., 2021) and some data augmentations (Brody et al., 2021) can increase the relation separability by changing the data format, these solutions still rely on the superficial features and fail to grasp the subtle semantic differences.

In this paper, we attempt to solve these two pre-

* Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2129–2140
October 12–17, 2022.

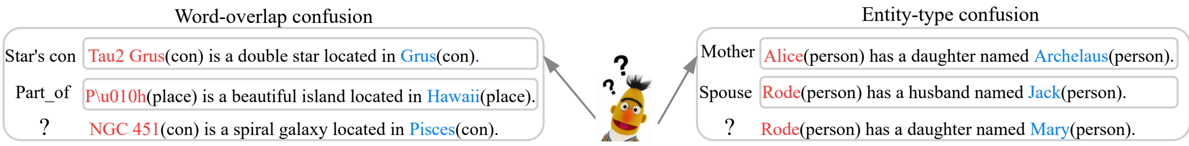


Figure 1: **Prediction confusions** in few-shot RE, including the word-overlap confusion (left) and the entity-type confusion (right). The word-overlap confusion has relations with the same grammatical structure and massive word overlaps but mismatched entity-type pair ("con-con" vs. "place-place"), while the entity-type confusion has relations with the same entity-type pair ("person-person" vs. "person-person"), where the head and tail entities are indicated in red and blue respectively and the word constellation is abbreviated to "con".

diction confusions by proposing a *discriminative rule-based knowledge* (DRK) method that consists of a logic-aware inference module and a discrimination finding module. Specifically, the logic-aware inference module relieves the word-overlap confusion by a rule-based incorporation of an ontology KG (Hao et al., 2019). Different from the simple concatenation in previous work , this module employs a logic rule (e.g., the "Star's con" relation happens when entity-type pair belongs to "con-con" rather than "place-place", shown in Fig.1) to constrain the model inference. In this way, the rule-based knowledge can not only mitigate the adverse effect of shallow text features but also provide a new inference direction. Unfortunately, this introduced rule is heavily dependent on the discrepancy of entity types, thereby losing its advantages in the entity-type confusion. To clear this confusion, the discrimination finding module is proposed, which employs a hierarchical contrastive learning strategy (the instance level and the category level) to further explore distinguishable text features.

We compare DRK with ML methods with and without external knowledge on four types of meta-tasks in FewRel 1.0 (Han et al., 2018). The results demonstrate: 1) DRK achieves significant improvements over the state-of-the-art baseline in terms of accuracy; 2) the logic-aware inference module is more effective than the discrimination finding one; 3) error analyses reveal that the word-overlap and entity-type confusions are the main error courses of mispredictions in few-shot RE, and DRK can effectively relieve these two prediction confusions.

Our key contributions are: 1) a *discriminative rule-based knowledge* (DRK) method for few-shot RE, which targets to relieve prediction confusions; 2) a logic-aware inference module for the word-overlap confusion by the rule-based knowledge incorporation, which opens a new inference direction for few-shot RE; 3) a discrimination finding module for the entity-type confusion by the hierarchical contrastive learning, which explores subtle semantic differences; 4) extensive experiments demonstrating the effectiveness of our proposals.

2 Related Work

2.1 Basic ML for Few-shot RE

In few-shot RE, basic ML generally infers relation type based on original text as the sole input, which can be roughly divided into two types, i.e., optimization-based ML and metric-based ML (Huang et al., 2021). Optimization-based ML focuses on finding good initialization points for parameters that can readily generalize to novel relation types within few gradient steps. For example, MAML adopts a model-agnostic gradient update strategy to produce good gradient stand points of parameters for novel relations (Finn et al., 2017). To reduce the computational complexity of MAML, Reptile only employs first-order derivatives to update parameters (Nichol et al., 2018). Metric-based ML aims to design a metric function that clearly measures the distance of instances in the embedding space. For instance, prototypical networks identify the relation labels by computing the similarity between query instances and the relation prototypes (Snell et al., 2017). Following this work, quantities of methods devote to improving the performance of prototype, e.g., Gao et al. (2019a) modify the representation of the prototype by highlighting the crucial instances and features, and Sun et al. (2019) redefine the prototype via a hierarchical attention scheme.

However, these ML methods can't make reliable inferences in the low-resource scenario, and easily fall into the prediction confusions. To clear these confusions, our proposal DRK introduces an ontology-level KG by the rule-based incorporation to avoid the adverse effect of shallow text inferences.

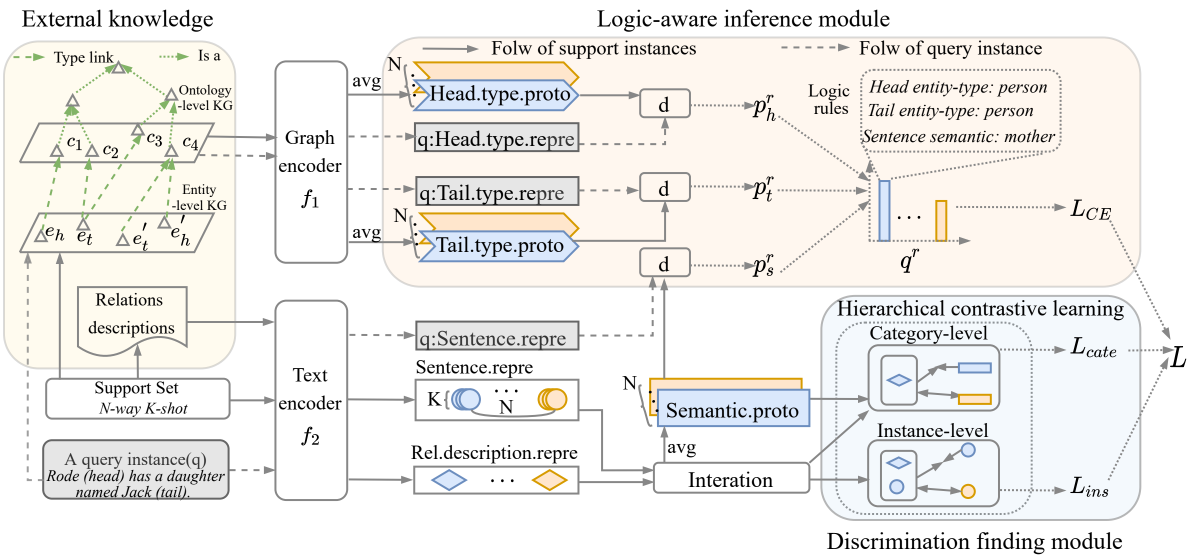


Figure 2: Framework of discriminative rule-based knowledge. For naming these process or submodules, "representation" and "prototype" are abbreviated as "repre" and "proto", respectively.

2.2 Knowledge-based ML for Few-shot RE

External knowledge has been widely employed in few-shot RE due to the abundant auxiliary semantic information. Based on the data structure, external knowledge can be divided into unstructured text and structured knowledge graph. For the unstructured text, most work focuses on leveraging relation and entity descriptions to enhance contextual semantic representation (Yang et al., 2020). Compared with the unstructured text, massive endeavors have been devoted to the structured knowledge graph. For example, Liu et al. (2020) inject entity-level triplets into text by a sentence tree, Roy and Pan (2021) integrate entity-level KG and text by several fusion techniques, and Yang et al. (2021) leverage an ontology-level knowledge graph to provide clues for the entity type and designed a fusion module based on self-attention to bridge the gap between the embeddings of text and the relation types. Besides, Sainz et al. (2021) annotate the entity type manually for each instance.

In essence, the above knowledge-based ML methods are still built on shallow inferences, suffering from prediction confusions. Our proposal adopts the logic-aware inference module and the discrimination finding module to clear the word-overlap and entity-type confusions, respectively.

3 Approaches

In this section, we first introduce the task definition as well as the framework of DRK in §3.1. Then, we detail the logic-aware inference module in §3.2 and the discrimination finding module in §3.3.

3.1 Task definition and model framework

RE. Formally, given a L_s-word instance s with the head and tail entities e_h and e_t, i.e., $s = \{w_1, \cdots, e_h, \cdots, e_t, \cdots, w_{L_s}\}$, the goal of RE is to correctly extract the relation triplet (e_h, e_t, r), where r is a relation label belonging to the predefined relation label set $\mathcal{R}$.

Few-shot RE. Following the typical N-way K-shot setting, a meta task consists of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. $\mathcal{S} = \{S_i\}_{i=1}^{N}$ has N novel relations $\mathcal{R} = \{r_i\}_{i=1}^{N}$, each relation r has $S_r = \{s_r^i\}_{i=1}^{K}$ containing K instances. Few-shot relation extraction targets to predict the relation label $r_Q \in \mathcal{R}$ of the query set $\mathcal{Q}$ based on the limited labeled data $\mathcal{S}$.

It is noteworthy that the few-shot RE models are trained on meta tasks sampled from instances of base relation labels then tested on meta tasks sampled from instances of novel relation labels(Zheng et al., 2021). And the base and novel relation labels are disjoint.

External Knowledge. Based on the data structure, external knowledge can be classified into the structured entity-level KG G_e, the structured ontology-level KG G_o and the unstructured knowledge G_t.

In specific, G_e consists of a set of relation triplets $\{(e_h, e_t, r) \in \mathcal{E} \times \mathcal{E} \times \mathcal{R}_e\}$, where $\mathcal{E}$ and $\mathcal{R}_e$ are the entity node set and the entity-level relation set, respectively. Similar to G_e, G_o can also be formulated as $\{(e_h, e_t, r) \in \mathcal{C} \times \mathcal{C} \times \mathcal{R}_o\}$, where $\mathcal{C}$, $\mathcal{R}_o$ are the ontology node set and the ontology-level relation label set, respectively. Note that there exhibits an "instance of" relation $\hat{r}$ between $\mathcal{E}$ and $\mathcal{C}$, i.e., $\{(e_h, e_t, \hat{r}) \in \mathcal{E} \times \mathcal{C} \times \hat{r}\}$. For unstructured knowledge G_t, it describes each relation r from the given relation set $\mathcal{R}$ with a L_r-word sequence relation description $a_r = \{w_i\}_{i=1}^{L_r}$.

Framework of DRK. As shown in Fig 2, In each iteration step, the support set $\mathcal{S}$ and the query set $\mathcal{Q}$ are first fed into external knowledge to obtain their corresponding structured and unstructured information that are encoded by the graph encoder f_1 and the text encoder f_2, respectively. And then, the logic-aware inference module employs the prototypes on structured and unstructured information to get the head, tail and contextual semantic representations, which are further constrained by the logic rule to compute L_{CE}. Furthermore, the discrimination finding module leverages the category-level contrastive learning L_{cate} and the instance-level contrastive learning L_{ins} to mine the distinguishable text features.

3.2 Logic-aware inference module

This module employs structured and unstructured knowledge encoders to get the entity-type and contextual semantic representations, which are further constrained by the logic rule.

3.2.1 Structured knowledge encoder

Each entity belongs a set of entity types, e.g., the entity "Biden" stemmed from the "politician" and "person" types. Take the head entity e_h for example, the ontology KG G_o[1] is adopted to construct it's type set C_h:

$$C_h = B(e_h), \tag{1}$$

where B is a link function base on the "instance_of" relation between e_h and c_h. Then a graph neural network encoder f_1 is employed to obtain the head entity type representation $\mathbf{H}_{e_h}$ of e_h:

$$\mathbf{H}_{e_h} = \frac{1}{|C_h|} \sum_{c \in C_h} f_1(c), \tag{2}$$

where $|C_h|$ is the number of elements in the type set. For any relation r, its support instances S_r contain K instances, hence the head entity type prototype of the relation r can be formulated as:

$$\mathbf{m}_h^r = \frac{1}{K} \sum_{e_h \in S_r} \mathbf{H}_{e_h}. \tag{3}$$

Analogously, the tail entity type prototype $\mathbf{m}_t^r$ of the relation r can also be obtained. And, for any query instance q in the query set $\mathcal{Q}$, similar encoding process can be employed to get its head and tail entity type representations: $\mathbf{H}_h^q$ and $\mathbf{H}_t^q$.

3.2.2 Unstructured knowledge encoder

Given a relation r with support instances $\{s_r^i\}_{i=1}^K$, the unstructured knowledge G_t [2] is retrieved to obtain the relation description a_r. Then, the support instances and the relation description are respectively fed into a text encoder $f_2(\cdot)$ to get the instance representations $\{\mathbf{H}_{s_r^i}^0 \in \mathbb{R}^{L_s \times d}\}_{i=1}^K$ and the relation-describing representations $\mathbf{H}_{a_r} \in \mathbb{R}^{L_r \times d}$, where d is the embedding dimension of f_2. To highlight the entities (Soares et al., 2019), the start tokens of the head and tail entities are concatenated to get the entity representations $\{\mathbf{H}_{s_r^i}^e \in \mathbb{R}^{1 \times 2d}\}_{i=1}^K$ that is different from Eq.(2).

Interaction. On the one hand, the relation description a_r elaborates the relation r, a_r can be utilized to further refine the instance representations. Taking the instance s_r^i as an example, its refined instance representation can be formulated as:

$$\begin{aligned} \mathbf{H}_{s_r^i}^{a_r} &= \sum_{j=1}^{L_s} \alpha_j \mathbf{H}_{s_r^i}^0[j:], \\ \boldsymbol{\alpha} &= soft\max\left(sum\left(\mathbf{H}_{s_r^i}^0 \mathbf{H}_{a_r}^T\right)\right), \end{aligned} \tag{4}$$

where $sum(\cdot)$ is a row-wise summation function, and hence the attention weight $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^{L_s}$ attends over the instance tokens. On the other hand, the relation description, as a highly-concise summary, also requires support instances to express its semantics. Specifically, for a specific relation r, its each support instance $\mathbf{H}_{s_r^i}^0$ attends its relation description a_r to obtain the instance-aware relation-describing representation $\mathbf{H}_{a_r}^{s_r^i}$:

$$\begin{aligned} \mathbf{H}_{a_r}^{s_r^i} &= \sum_{j=1}^{L_r} \beta_j \mathbf{H}_{a_r}[j:], \\ \boldsymbol{\beta} &= soft\max\left(sum\left(\mathbf{H}_{a_r}\left(\mathbf{H}_{s_r^i}^0\right)^T\right)\right), \end{aligned} \tag{5}$$

[1] The construct process of G_o can refer to https://github.com/imJiawen/KEFDA

[2] All the relation descriptions in G_t is shown in Table 1 of Appendix A.1

where $\boldsymbol{\beta} = \{\beta_j\}_{j=1}^{L_r}$ is the attention weight over the relation description tokens and $\mathbf{H}_{a_r}^{s_r^i}$ elaborates the instantiation traits of the relation r.

Contextual semantic representation. After the interaction between the instance text and the relation description, the final contextual semantic representation can be further formulated as:

$$\mathbf{H}_{s_r^i} = [\mathbf{H}_{s_r^i}^e; \mathbf{H}_{s_r^i}^{a_r}] + Mul(\mathbf{H}_{a_r}^{s_r^i}), \tag{6}$$

where $\mathbf{H}_{s_r^i}^e \in \mathbb{R}^{1\times 2d}$, $\mathbf{H}_{a_r}^{s_r^i} \in \mathbb{R}^{1\times d}$, $\mathbf{H}_{s_r^i}^{a_r} \in \mathbb{R}^{1\times d}$ and $\mathbf{H}_{s_r^i} \in \mathbb{R}^{1\times 3d}$. $Mul(\cdot)$ is a multi-layer perceptron that converts the dimension from d to 3d.

Similar to Eq.(3), the contextual semantic prototype of r can be obtained:

$$\mathbf{m}_s^r = \frac{1}{K}\sum_{i=1}^{K} \mathbf{H}_{s_r^i}. \tag{7}$$

Note that any instance q in the query set $\mathcal{Q}$ does not have the relation description. Therefore, the refined instance representations in Eq.(4) is replaced with the average pooling and the $Mul(\cdot)$ operation in Eq.(6) is removed when defining the contextual semantic representation of q, i.e., $\mathbf{H}_s^q$.

3.2.3 Logic rule

For RE, this logic rule is assumed as a set of conditions that should be occurred simultaneously (Han et al., 2021a). Taking the "Mother" relation in Fig.1 as an example, whether an instance expresses this relation must satisfy three conditions: 1) the head entity is the person type; 2) the tail entity is the person type; 3) the contextual semantics of this instance describes the "Mother" relation. Based on this logic rule, such three conditions can correspondingly be transformed into three probabilities.

Specifically, for a query instance q, the probability of the head entity type belonging to relation r is:

$$p_h^r = \frac{\exp\left(d\left(\mathbf{H}_h^q, \mathbf{m}_h^r\right)\right)}{\sum\limits_{n=1}^{N} \exp\left(d\left(\mathbf{H}_h^q, \mathbf{m}_h^n\right)\right)}, \tag{8}$$

where $d(\cdot,\cdot)$ indicates the dot product. Analogous to Eq.(8), the probabilities of the tail entity and the contextual semantics belonging to the relation r can also be obtained, i.e., p_t^r and p_s^r. With these three probabilities, the final probability of q expressing the relation r is calculated as:

$$q^r = p_h^r \cdot p_t^r \cdot p_s^r. \tag{9}$$

Consequently, the cross entropy loss is used to optimize the RE parameters:

$$L_{CE} = -\sum_{q\in\mathcal{Q}} \log(q^r), \tag{10}$$

where the ground-truth relation label of the query instance q is r.

3.3 Discrimination finding module

Despite effective in the word-overlap confusion, the logic-aware inference module cannot handle the entity-type confusion with the matched entity-type pair. Then we propose a hierarchical contrastive learning (including the the instance level and the category level) to mine subtle differences of relation instances.

3.3.1 Instance-level contrastive learning

In a support set, we formulate the instance-level contrastive (He et al., 2020) as follows:

$$L_{\text{ins}} = \frac{-1}{NK^2}\sum_{r=1}^{N}\sum_{i=1}^{K} log\frac{\sum_{j=1}^{K} exp(\mathbf{H}_{s_r^i}\cdot\mathbf{H}_{s_r^j}/\tau_1)}{\sum\limits_{r'\neq r}\sum_{k=1}^{K}(\mathbf{H}_{s_r^i}\cdot\mathbf{H}_{s_{r'}^k}/\tau_1)}, \tag{11}$$

where τ_1 is a temperature hyperparameter, $\mathbf{H}_{s_r^i}$ and $\mathbf{H}_{s_r^j}$ belonging to the same relation form positives and L_{ins} aims to pull the positives closer.

3.3.2 Category-level contrastive learning.

Similar to Eq.(3), the instance-aware relation-describing prototype of r can be obtained, i.e.,

$$\mathbf{m}_a^r = \frac{1}{K}\sum_{i=1}^{K}\mathbf{H}_{a_r}^{s_r^i}. \tag{12}$$

This prototype $\mathbf{m}_a^r$ summarizes the instance-aware traits of relation r. When encountering the prediction confusion, these traits can help RE models to distinguish the subtle differences of instances.

In a support set, the contextual semantic prototype $\mathbf{m}_s^r$ must be close to the belonged relation-describing prototype $\mathbf{m}_a^r$, and keep away from the others. Hence, the category-level contrastive learning is formulated as:

$$L_{cate} = \frac{-1}{N}\sum_{r=1}^{N}\log\frac{\exp\left(\mathbf{m}_s^r\cdot\mathbf{m}_a^r/\tau_2\right)}{\sum\limits_{j\neq r}^{N}\exp\left(\mathbf{m}_s^r\cdot\mathbf{m}_a^j/\tau_2\right)}, \tag{13}$$

where τ_2 is a temperature hyperparameter.

In all, the overall training objective is:

$$L = L_{CE} + L_{ins} + L_{cate}, \tag{14}$$

Task	#Rel	#Ins	Len	Link(%)
Training	50	35000	25	98.35
validation	14	9800	24	98.46
testing	16	11200	24	98.70

Table 1: Statistics of FewRel 1.0. "#Rel" and "#Ins" denote the number of relations and instances, respectively. "Len" means the average token length of instances. "Link" demonstrates the probability of an entity linking to a corresponding type in G_o.

4 Experiments

4.1 Dataset

Experiments are conducted on FewRel 1.0 [3] (Han et al., 2018) that consists of 100 relations extracted from Wikipedia. Since the test set with 20 relations is unpublished, following previous work (Yang et al., 2020, 2021), we re-split the published 80 relations into 50, 14 and 16 for training, validation and testing, respectively. The statistics of the re-splitted dataset can refer to Tabel 1. The ontology-level KG we adopt in this paper comes from KEFDA [4]. Besides, the unstructured text knowledge for test data is represented in Table 5 in Appendix A.1.

4.2 Model Configuration

The model configurations are kept the same among all discussed models, including our proposal and the selected baselines. In detail, following (Han et al., 2018, 2021b), we use the classification accuracy to evaluate the performance of DRK on four typical meta tasks (Han et al., 2018): 5-way 1-shot, 5-way 5-shot, 10-way 1-shot and 10-way 5-shot. We use BERT$_{base}$ (Soares et al., 2019) as the text feature encoder f_2 and apply DistMult (Yang et al., 2015) as the graph neural network (Zheng et al., 2020b) encoder f_1. The embedding dimension of f_2 and f_1 is 768 and 256, respectively. We train and test each model with 20,000 and 10,000 iteration steps, respectively. Besides, we set the batch size to 4, the weight decay to 2×10^{-5}, the max sentence length L_s to 128, τ_1 to 1, and τ_2 to 1.

4.3 Baselines

As mentioned above, we introduce several basic ML and knowledge-based ML methods as baselines. In specific, the basic ML approaches contain **GNN** (Satorras and Estrach) that considers all the instances in a meta task as nodes in a graph, then leverages the label propagation to infer relations, **Snail** (Mishra et al., 2018) that combines temporal convolutions and soft attention to learn information from past experiences to predict relations, **Siamese** (Mishra et al.) that employs siamese neural networks to distinguish relations, **Proto** (Snell et al., 2017) that calculates the similarity between the query instance and the relation prototypes to predict relation types, and **BERT-PAIR** (Gao et al., 2019b) that concatenates the query instances with all labeled support instances as pair sequences, then identifies the relation type through the similarity in each pair. ; while the knowledge-based ML approaches include **ConceptFERE** [5] (Yang et al., 2021) that concatenates the entity type with sentence text by a self-attention fusion, and **KEFDA** (Zhang et al., 2021) that also concatenates the entity type and sentence text, and leverages a relation-meta learning network (Chen et al., 2019) to learn implicit relation matching. Note that all the baselines and our proposal use the same text feature and graph encoders for fair comparisons.

4.4 Overall Evaluation

To examine the few-shot relation extraction performance of our proposal as well as the baselines, we report the overall evaluation results on four types of meta tasks in Table 2.

Among all baselines, the knowledge-based ML ConceptFERE achieves the best performance with the average 1.11% improvement against BERT-PAIR (the best basic ML) on the four meta tasks. This performance advancement reflects the benefits brought by the external knowledge that provides auxiliary clues for relation extraction. Compared with ConceptFERE, our proposal DRK can further enhance the extraction performance with the respective improvements of 6.72%, 2.30%, 10.73% and 4.17% on the four meta task. Such model improvements prove the effectiveness of DRK by relieving the prediction confusions.

Clearly, as the shot number decreases, all discussed models get impaired. For example, comparing meta tasks on 5-way 1-shot and 5-way 5-shot, the best baseline ConceptFERE loses the relative 6.71% model performance (9.56% in 10-way 1-shot vs. 10-way 5-shot). These phenomena show that these models are sensitive to the shot number,

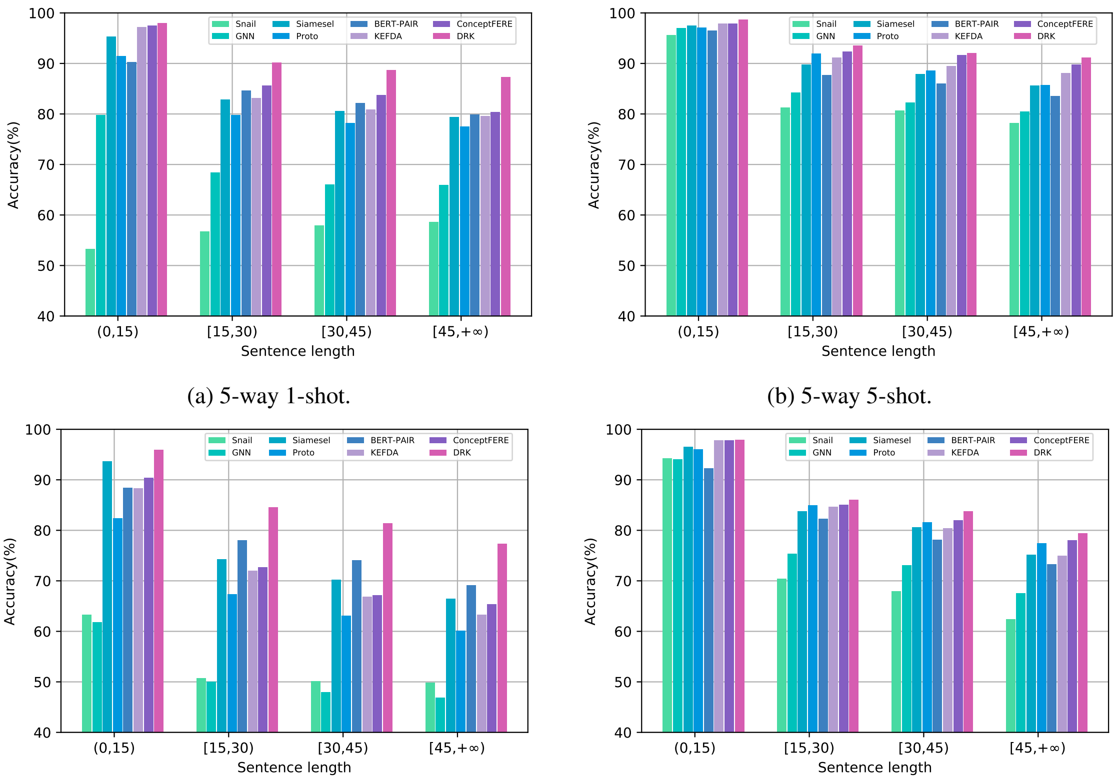
[3]https://github.com/thunlp/FewRel

[4]https://github.com/imJiawen/KEFDA

[5]We use the ConceptFERE(simple) version here to allow for the computation overheads in the training of "5-way 5-shot" and "10-way 5-shot" meta tasks.

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Snail	57.82	80.53	50.40	68.11
GNN	66.48	82.65	48.14	73.22
Siamese	81.29	88.18	71.00	81.12
Proto	78.59	88.99	64.07	81.80
BERT-PAIR	82.57	89.00	73.37	81.81
KEFDA	80.46	89.88	68.23	81.49
ConceptFERE†	<u>84.28</u>	<u>90.34</u>	<u>74.00</u>	<u>81.82</u>
DRK(Our)	**89.94**$^{\triangle}$	**92.42**$^{\triangle}$	**81.94**$^{\triangle}$	**85.23**$^{\triangle}$

Table 2: Few-shot relation extraction performance in terms of accuracy(%) on four types of meta tasks. † means the results are quoted from the original paper (Yang et al., 2021). Results of the best baseline and the best performer in each column are <u>underlined</u> and **boldfaced**, respectively. Statistical significance of pairwise differences of DRK vs. the best baseline is determined by a t-test ($^{\triangle}$ for $\alpha = 0.05$).



(a) 5-way 1-shot.

(b) 5-way 5-shot.

(c) 10-way 1-shot.

(d) 10-way 5-shot.

Figure 3: Effect on the performance of our proposal and baselines affected by the sentence length.

and prone to fail when the labeled data become fewer. In the same comparisons about the shot number, however, DRK only has 2.68% and 3.86% performance wastage, indicating that DRK is more robust than other methods in the low-data regime.

4.5 Impact of sentence length

To explore the performance of few-shot RE affected by the sentence length, we group the testing performance of our proposal and baselines based on the sentence length L_s. Considering the distribution of testing data, the sentence length is divided into four groups, i.e., $L_s \in (0, 15), [15, 30), [30, 45), [45, +\infty)$. With these sentence groups, the results on the 5-way 1-shot, 5-way 5-shot, 10-way 1-shot and 10-way 5-shot meta task are plotted in Fig.3. Then, we take the performance on the 5-way 1-shot meta task for instance to analyze the results.

In general, with the increase of sentence length,

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
DRK	89.94	92.42	81.94	85.23
- logic&text	84.08	90.48	79.02	81.90
- logic&type	83.64↓	90.37↓	74.60↓	81.45↓
- instance	88.07	91.80	80.53	83.43
- category	85.22	90.63	79.83	81.63

Table 3: Ablation studies on DRK, where "-logic&text" ("-logic&type") means that the logic-aware inference module is replaced with prototypical networks that do not use entity type from ontology-level KG (the simple concatenation of incorporating entity type from ontology KG), while "-instance" ("-category") denotes the removal of the instance-level (category-level) contrastive learning, and the biggest drop in each column is appended ↓.

almost all models (except for Snail on 5-way 1-shot meta task in Fig.3a) exhibit a ever-decreasing trend in terms of the model performance. This downward trend can be attributed to the fact that short sentences are more concise than long sentences in the grammatic structure, and hence easier for RE models to understand their semantics. In particular, the steep performance degenerations in KEFDA and ConceptFERE shown in Fig.3a further confirm our hypothesis about the existing knowledge-based ML methods. Furthermore, the upward trend of Snail in Fig.3a may be due to the probable error in the low-shot setting. In other three meta tasks, Snail still keeps the decreasing trend.

Next, let's focus on our proposal DRK. Clearly, DRK still presents notable performance advantages over all baselines for each sentence group. For example, compared to the best baseline ConceptFERE on 5-way 1-shot meta task in Fig.3a, DRK achieves 0.47%, 4.51%, 4.90% and 6.96% model improvements in terms of accuracy when $L_s \in (0, 15), [15, 30), [30, 45), [45, +\infty)$, respectively. Interestingly, the improvement magnitude of DRK always keeps pace with the increase of sentence length. This increasing pattern can be explain by the fact that DRK can effectively extract distinguishable features from overlong sentences, thereby reducing the adverse effect of massive noises in sentences.

Similar findings can also be found on the 5-way 5-shot, 10-way 1-shot and 10-way 5-shot meta tasks. The ever-decreasing trend in the 5-way 1-shot also appears in these meta tasks. Comparing 1-shot and 5-shot meta tasks, all the methods achieve improvements for four sentence groups due to the increase of training data. Besides, compared to baselines, our proposal achieves the biggest improvements for four sentence groups in Fig.3c than on the other three meta tasks. This phenomenon shows that our proposal is robust and competitive on the most challenging 10-way 1-shot meta task.

4.6 Ablation Studies

In order to better understand the contributions of different components, the ablation studies are conducted on four types of meta tasks. In the ablation studies, we remove or replace some specific parts to measure their influence on DKR, which is marked with the notation"-". The ablation results are demonstrated in Table 3.

Clearly, the removal or replacement of components all leads to the model degeneration, proving the efficacy of each component. In particular, the biggest drop happens in "-logic&type", which reflects that the logic-aware inference module plays a key role in few-shot RE and simple concatenation in previous works probably cannot improve the model performance. These findings can be further verified in the comparison of "-logic&text" and "-logic&type", where the model incorporating the ontology-level KG ("-logic&type") loses the competitions against the model only using text ("-logic&text"). Besides, in the discrimination finding module, losing the category-level contrastive learning is more sever than DRK without the instance level one. The comparisons demonstrate that compared to the instance-level contrastive learning, the category-level one can better identify inter-class differences to help predict relations.

4.7 Error analyses

To further figure out the error causes, we conduct error analyses of our proposal and the best baseline ConceptFERE on the test set. In particular, we divide the causes of classification errors into there main categories: word-overlap (word), entity-type (entity) and others. For each test relation, we randomly select 5 instances (i.e., K=5) and 50

Relation	Correct	Word	Entity	Other	Top 2 confounders
constellation	0.98	0.02	0.00	0.00	follows(0.02)
	0.96	0.03	0.00	0.01	part of(0.02),located in body of water(0.02)
part of	0.50	0.35	0.08	0.07	follows(0.10)
	0.40	0.39	0.11	0.10	member(0.18), subject(0.14)
mother	0.92	0.00	0.08	0.00	child(0.06), spouse(0.02)
	0.52	0.00	0.48	0.00	child(0.42), spouse(0.06)
spouse	0.48	0.00	0.52	0.00	child(0.44), mother(0.08)
	0.36	0.02	0.72	0.00	child(0.34), mother(0.28)

Table 4: Rates of the correct predictions (Correct), the word-overlap error (Word), the entity-type error (Entity) and the other errors (Other) in DRK (shown in orange) and ConceptFERE (shown in blue). Top 2 confounders list the top two wrongly-predicted relations and their rates.

instances of this relation into the support set and the query set, respectively. And in the evaluating process, the support set is used to calculate the relation prototype, then predict the relation label in the query set based on the similarity to the relation prototype. Also, we repeat the process three times and reported the mean result. Note that the prediction rate in the query set is presented relation-by-relation. Based on this setting, the error analyses of some relations are listed in table 4, and the other test relations can refer to Appendix A.2.

As shown in table 4, the rate of correct prediction varies widely among different relations, especially for some confusion relations ("part of" and "spouse"), their performance are much lower than expected. This unexpected performance is stemmed from the "word" and "entity" error types. Besides, the top 2 confounders have similar expressions or the same entity-type pairs as the instances of the ground-truth label. These findings all verify our arguments that owing to the limited inference based on the shallow features, RE models easily fall into the prediction confusions.

Fortunately, DRK relieves the prediction confusion by reducing the rates of these two errors. Taking "spouse" as an example, DRK achieves obvious declines in the "word" and "entity" errors against ConceptFERE (reducing 0.02 and 0.20 rates, respectively). Furthermore, DRK lowers the number of confounders relations. For example, taking "constellation" for example, ConceptFERE has two confounders ("part of" and "located in body of water") while DRK only has one confounder ("follows"). The above phenomena demonstrate that our proposal can further boost RE performance for each relation by reducing these errors, rather than improving easy relations for good average performance as the baselines do.

5 Conclusion

This paper focuses on the prediction confusions in few-shot RE. To relieve these confusions, this paper develops a *discriminative rule-based knowledge* (DRK) method consisting of a logic-aware inference module and a discrimination finding module. Specifically, the first module relieves the word-overlap confusion through the rule-based knowledge incorporation and the other module alleviates the entity-type confusion by a hierarchical contrastive learning. Extensive experiments show the effectiveness of our proposal. As for future work, we plan to explore the confusion by the parameter-efficient prompt tuning (Lester et al., 2021; Liao et al., 2022).

References

Sam Brody, Sichao Wu, and Adrian Benton. 2021. Towards realistic few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Mingyang Chen, Wen Zhang, Wei Zhang, Qiang Chen, and Huajun Chen. 2019. Meta relational learning for few-shot link prediction in knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics*, pages 45–57.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019a. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence*.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019b. Fewrel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021a. PTR: prompt tuning with rules for text classification. *CoRR*, abs/2105.11259.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yi Han, Linbo Qiao, Jianming Zheng, Zhigang Kan, Linhui Feng, Yifu Gao, Yu Tang, Qi Zhai, Dongsheng Li, and Xiangke Liao. 2021b. Multi-view interaction learning for few-shot relation classification. In *The 30th ACM International Conference on Information and Knowledge Management*.

Junheng Hao, Muhao Chen, Wenchao Yu, Yizhou Sun, and Wei Wang. 2019. Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1709–1719.

Keqing He, Jinchao Zhang, Yuanmeng Yan, Weiran Xu, Cheng Niu, and Jie Zhou. 2020. Contrastive zero-shot learning for cross-domain slot filling with adversarial attack. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Weizhi Huang, Ming He, and Yongle Wang. 2021. A survey on meta-learning based few-shot classification. In *Machine Learning and Intelligent Communications - 6th EAI International Conference*, volume 438, pages 243–253.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Luoqiu Li, Xiang Chen, Hongbin Ye, Zhen Bi, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. On robustness and bias analysis of bert-based relation extraction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction - 6th China Conference*.

Jinzhi Liao, Xiang Zhao, Jianming Zheng, Xinyi Li, Fei Cai, and Jiuyang Tang. 2022. PTAU: prompt tuning for attributing unanswerable questions. In *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *6th International Conference on Learning Representations*.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A simple neural attentive meta-learner. In *6th International Conference on Learning Representations*.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999.

Meng Qu, Tianyu Gao, Louis-Pascal A. C. Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7867–7876.

Arpita Roy and Shimei Pan. 2021. Incorporating medical knowledge in BERT for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing.

Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In 6th International Conference on Learning Representations.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, pages 4077–4087.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In Proceedings of the 57th Conference of the Association for Computational Linguistics.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.

Prasetya Ajie Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. 2021. Avoiding inference heuristics in few-shot prompt-based finetuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

Hailin Wang, Guoming Lu, Jin Yin, and Ke Qin. 2021. Relation extraction: A brief survey on deep neural network based methods. In The 4th International Conference on Software Engineering and Information Management.

Yingyao Wang, Junwei Bao, Guangyi Liu, Youzheng Wu, Xiaodong He, Bowen Zhou, and Tiejun Zhao. 2020. Learning to decouple relations: Few-shot relation classification with entity-guided attention and confusion-aware training. In Proceedings of the 28th International Conference on Computational Linguistics.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In 3rd International Conference on Learning Representations.

Kaijia Yang, Nantao Zheng, Xinyu Dai, Liang He, Shujian Huang, and Jiajun Chen. 2020. Enhance prototypical network with text descriptions for few-shot relation classification. In The 29th ACM International Conference on Information and Knowledge Management.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. Entity concept-enhanced few-shot relation extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 987–991.

Jiawen Zhang, Jiaqi Zhu, Yi Yang, Wandong Shi, Congcong Zhang, and Hongan Wang. 2021. Knowledge-enhanced domain adaptation in few-shot relation classification. In The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

Jianming Zheng, Fei Cai, and Honghui Chen. 2020a. Incorporating scenario knowledge into A unified fine-tuning architecture for event representation. In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval.

Jianming Zheng, Fei Cai, Wanyu Chen, Wengqiang Lei, and Honghui Chen. 2021. Taxonomy-aware learning for few-shot event detection. In The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, pages 3546–3557.

Jianming Zheng, Fei Cai, Yanxiang Ling, and Honghui Chen. 2020b. Heterogeneous graph neural networks to predict what happen next. In Proceedings of the 28th International Conference on Computational Linguistics.

A Appendix

A.1 Dataset

We demonstrate some unstructured text knowledge, including the relation id,, name and description in Table 5.

A.2 Errors in test relations

Following the setting in section §4.7, the detail prediction performance of DRK and ConceptFERE on each test relation is shown in Table 6.

In general, our proposal outperforms the baseline for almost all test relations, and the observations from section §4.7 can also be found in Table 6. Besides, compared to the ConceptFERE that only precisely identifies one relation (P2094), our proposal can accurately identify four relations (P177, 2094, P412 and P413), which demonstrates our proposal has a competitive inference ability. However, our proposal underperforms the baseline for the relation P206. This unusual relation may be stemmed from an improper representation of entity in the linking process, thereby invalidating our proposed modules.

Id	Ralation name	Relation description
P155	follows	immediately prior item in a series of which the subject is a part.
P177	crosses	obstacle (body of water, road, ...) which this bridge crosses over or this tunnel goes under.
P206	located in or next to body of water	located in or next to body of water", "sea, lake or river".
P2094	competition class	official classification by a regulating body under which the subject (events, teams, participants, or equipment) qualifies for inclusion.
P25	mother	female parent of the subject. For stepmother, use stepparent.
P26	spouse	the subject has the object as their spouse (husband, wife, partner, etc.)
P361	part of	object of which the subject is a part (it's not useful to link objects which are themselves parts of other objects already listed as parts of the subject).
P364	original language of film or TV show	language in which a film or a performance work was originally created.
P40	child	subject has object as biological, foster, and/or adoptive child.
P410	military rank	military rank achieved by a person.
P412	voice type	person's voice type. expected values: soprano, mezzo-soprano, contralto, countertenor, tenor, baritone, bass (and derivatives).
P413	position played on team	position or specialism of a player on a team, e.g. Small Forward.
P463	member of	organization or club to which the subject belongs. Do not use for membership in ethnic or social groups, nor for holding a position such as a member of parliament.
P59	constellation	the area of the celestial sphere of which the subject is a part (from a scientific standpoint, not an astrological one).
P641	sport	sport in which the subject participates or belongs to.
P921	main subject	primary topic of a work.

Table 5: Relation descriptions for the test set with the relation id, name and description content.

Id	Correct		Top 2 confounders	
P155	0.88	0.86	P361(0.04), P463(0.04)	P361(0.04), P463(0.04)
P177	1.00	0.96	None	P361(0.02), P206(0.02)
P206	0.42	0.62	P177(0.38),P361(0.20)	P177(0.38)
P2094	1.00	1.00	None	None
P25	0.92	0.52	P40(0.06), P26(0.02)	P40(0.42), P26(0.06)
P26	0.48	0.36	P40(0.44), P25(0.08)	P25(0.34), P26(0.28)
P361	0.50	0.40	P463(0.20), P155(0.10)	P463(0.18), P921(0.14)
P364	1.00	0.98	None	P921(0.02)
P40	0.82	0.66	P26(0.06), P25(0.06)	P25(0.28), P26(0.06)
P410	0.98	0.98	P413(0.02)	P413(0.02)
P412	1.00	0.96	None	P413(0.04)
P413	1.00	0.82	None	P412(0.14), P2094(0.04)
P463	0.82	0.76	P361(0.10), P412(0.06)	P361(0.08), P26(0.04)
P59	0.98	0.96	P155(0.02)	P206(0.02), P361(0.02)
P641	0.48	0.82	P413(0.30), P2094(0.10)	P412(0.16), P364(0.02)
P921	0.52	0.40	P361(0.24), P641(0.06)	P463(0.18), P59(0.14)

Table 6: Rates of the correct prediction and top 2 confounders in our proposal DRK (shown in orange) and ConceptFERE (shown in blue)