mathematics

MDPI

Article

Taxonomy-Aware Prototypical Network for Few-Shot Relation Extraction

Mengru Wang, Jianming Zheng and Honghui Chen *

Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China
* Correspondence: chenhonghui@nudt.edu.cn

Abstract: Relation extraction aims to predict the relation triple between the tail entity and head entity in a given text. A large body of works adopt meta-learning to address the few-shot issue faced by relation extraction, where each relation category only contains few labeled data for demonstration. Despite promising results achieved by existing meta-learning methods, these methods still struggle to distinguish the subtle differences between different relations with similar expressions. We argue this is largely owing to that these methods cannot capture unbiased and discriminative features in the very few-shot scenario. For alleviating the above problems, we propose a taxonomy-aware prototype network, which consists of a category-aware calibration module and a task-aware training strategy module. The former implicitly and explicitly calibrates the representation of prototype to become sufficiently unbiased and discriminative. The latter balances the weight between easy and hard instances, which enables our proposal to focus on data with more information during the training stage. Finally, comprehensive experiments are conducted on four typical meta tasks. Furthermore, our proposal presents superiority over the competitive baselines with an improvement of 3.30% in terms of average accuracy.

Keywords: few-shot; relation extraction; prototype; distribution calibration; contrastive learning

MSC: 68T50

Citation: Wang, M.; Zheng, J.; Chen, H. Taxonomy-Aware Prototypical Network for Few-Shot Relation Extraction. *Mathematics* **2022**, *10*, 4378. https://doi.org/10.3390/math10224378

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 3 October 2022
Accepted: 15 November 2022
Published: 21 November 2022

1. Introduction

Relation extraction (RE) is designed to extract the relation between two entities in a given text [1], and has been widely applied in downstream tasks of Nature Language Processing, e.g., knowledge base population and question answering [2]. Traditional deep neural network methods [3] for RE are typically challenged by the need to gather large amounts of high-quality annotation data, which is expensive and laborious. Therefore, few-shot relation extraction is feasible for realistic applications [4]. Furthermore, meta-learning methods are proposed to address such a low-resource dilemma [5]. The core of meta-learning (ML) is to optimize methods via diverse meta-tasks, each with several labeled instances, so that the methods can rapidly learn to identify new relations with only few instances. Figure 1 illustrates an instance of two-way one-shot for few-shot RE.

These ML approaches can be broadly classified into three categories, namely model-, optimization- and metric-based ML methods [6]. As a popular solution, the metric-based ML methods focus on designing a metric function in order to identify the distance between instances in the query set and the categories (illustrated with a few instances) appearing in the support set. Prototypical network [7], a simple and effective metric-based ML method, approximately represents each category via a prototype, which is achieved through averaging the embeddings of these instances that belong to the class. A great deal of works are devoted to improving the representation of prototypes, e.g., Gao et al. [8] modifies the representation of prototypes by highlighting the crucial instances and features, and Wen et al. [5] integrates the transformer model into prototype nets for greater expressiveness. In addition,

some recent works have utilized external knowledge to provide more clues to the representation of prototypes, e.g., Qu et al. [9] optimizes the posterior distribution of a prototype via a global relation graph as the initial prior of the prototype, Yang et al. [10] employs the text descriptions of relations and entities to enhance representations of a prototype, and Yang et al. [11] fuses the entity concept to constrain the representations of a prototype.

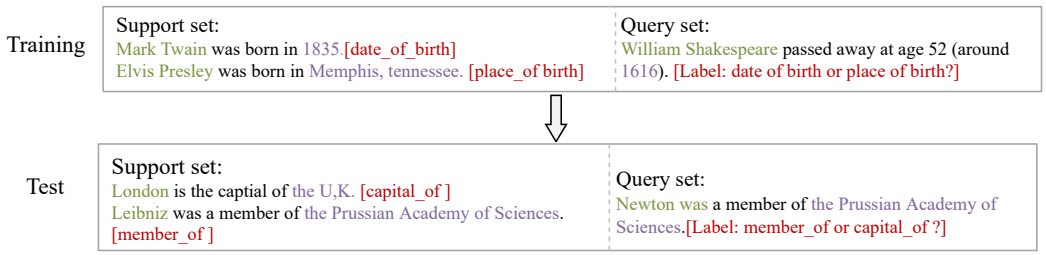


Figure 1. A 2-way 1-shot meta task for few-shot relation, where the head entity, the tail entity and the relation category are in green, purple and red, respectively. The relation categories in training and test stage are disjoint.

However, there are two main limitations to these methods. First, the prototype representation of the above models bears some bias and the discriminative ability is insufficient in few-shot scenarios, which restricts the performance of these ML methods. Additionally, these improved methods usually design complex structures and introduce excessive parameters, which increases the computational burden and also easily leads to overfitting in the few-shot schema. Second, current ML methods treat all training instances equally [6] or pay more attention to very hard instances [12,13], which prevents these methods from extracting useful information from the training instances. Intuitively, on the one hand, tasks that are overly simple provide no valuable information; on the other hand, even humans can only extract critical information from moderately hard instances and struggle with very hard instances, let alone neural network models.

With the aim of alleviating the above problems, we propose a taxonomy-aware prototypical network (TAPN) method, consisting of two modules: a category-aware calibration module and a task-aware training strategy module. Specifically, the category-aware calibration module leverages relation description to explicitly calibrate the prototype distribution in order to obtain unbiased representations and applies prototype-aware contrastive learning to implicitly calibrate the prototype representations to be more discriminative. The task-aware training strategy module leverages the task-aware difficulty to balance the weights of easy and hard instances, which also dynamically adapt different meta tasks.

We evaluate our proposal on four classic meta tasks, and the broad results of the experiment indicate that TAPN is markedly superior to baselines. Additionally, ablation research further validates the effectiveness of these two modules and an error analysis shows the interpretability of TAPN's good performance.

In summary, our major contributions can be summarized as follows:

(1) To the best of our knowledge, we are the first to explicitly and implicitly calibrate the prototype representation simultaneously without introducing extra or even harmful parameters.
(2) We design a category-aware calibration module to enable the representation of an unbiased and more discriminative prototype by relation description and prototype-aware contrastive learning, respectively.
(3) We propose a task-aware training strategy module to extract beneficial knowledge by exploring hard task and sample instances.
(4) The experimental findings confirm the validity of our model in terms of accuracy against the competing baselines.

The remainder of the paper is organized as follows: We review the related work for few-shot RE in Section 2, detail our approach in Section 3, design our experiments

in Section 4, analyze the results of our proposal in Section 5 and conclude our work in Section 6.

2. Related Work

2.1. Relation Extraction

RE is designed to determine the relations between entities in a given sentence. Most traditional RE models extract the relations under supervised settings [14], which can be classified into three categories: neural-, kernel- and feature-based methods [1].

Feature-based methods [15] typically focus on generating a set of features, e.g., word feature [16], syntactic feature [17], semantic feature [18], etc., for the relation classifier. Kernel-based methods mainly design the kernel functions to compute similarities between two relation instances. These kernel functions comprise syntactic sequence kernel [19], tree kernel [20], dependency tree kernel [21], dependency graph path kernel [22] and composite kernels [23]. Distinct from manual participation in the feature-based or kernel-based methods, neural-based models [1] often concentrate on extracting the relational features with neural networks, e.g., convolution neural network (CNN) [24], graph convolutional network [25,26], and recurrent neural network [27], to perform the end-to-end training.

Typically, the aforementioned approaches work well based on numerous labeled data. However, it is time-consuming [9] and impractical to collect such massive annotated data in some professional domains. We focus on extracting relation triple in the few-shot scenario.

2.2. Few-Shot Relation Extraction

Meta-learning methods [28] have been extensively applied to the few-shot RE. The ML models are trained in various meta-tasks with few instances as demonstrations, then can be generalized to new meta tasks. In general, these ML methods are divided into three category [29]: metric-, optimization- and model-based methods [17].

Model-based methods [30] emphasize on designing the architecture of the model to address the few-shot task. To be specific, MANN [31] designs a memory-enhanced neural network to quickly absorb new data and proposes an effective strategy for accessing the external memory, which provides the ability to quickly predict new relations. Optimization-based methods [32,33] try to initialize the parameters well. For instance, Finn et al. [32] optimize parameters with few training data so that they can be adapted to novel tasks with a limited number of gradient descent steps. The metric-based approaches focus on learning a metric function to determine the similarity between support sentences and query sentences. For instance, relation networks [34] learn a deep distance metric on the basis of the neural network instead of the fixed Euclidean distance or dot product. The prototypical network [7] predicts relation labels through computing the similarity between the prototype of each class and query sentences, which is derived from averaging the representations of all the examples belonging to a particular class. In addition, a great deal of works are designed to improve the prototypical network: Gao et al. [8] present hybrid attention-based prototypical networks to deal with the diversity and noise of text, Han et al. [12] introduce external relation description and combine global and local features as hybrid prototypes, that learns better characterization through utilizing relational label information.

However, these improved prototypical networks almost introduce extra parameters, e.g., parameters of the attention mechanism, which require sufficient data for optimization and is not realistic in the few-shot scenario. In addition, the prototype representations are usually biased and insufficiently discriminative. In this paper, compared to vanilla prototypical networks, we calibrate the prototype representation without introducing additional parameters.

2.3. Contrastive Learning

Contrastive Learning achieves success in computer vision (CV) [35] through pulling together positive instances and pushing negative instances away simultaneously. Different from positives produced by cropping, flipping, distortion and rotation in CV, methods to

construct positives for discrete text sequences present a critical problem. Moreover, there are quantities of works dedicated to solving the above problem. For example, to design proper positives, Wu et al. [36] and Meng et al. [37] design word deletion, reordering and substitution techniques, Yan et al. [38] propose four new data-augmentation techniques (adversarial attack, token shuffling, cutoff and dropout), Gao et al. [39] apply random dropout as noise for sentence text and Jiang et al. [40] introduce different templates to express the same sentence text.

However, the aforementioned works usually construct positives and negatives at the instance-level and ignore the connection between instances and categories. Inspired by this, we design a prototype-aware contrastive learning prototype at the category-level, which drives the representations of categories to become more discriminating.

3. Approaches

In this section, we will display the details of our proposed technique. As exhibited in Figure 2, the structure of our proposal includes two modules: the category-aware calibration module and the task-aware training strategy module. In detail, the feature-encoder first transforms the input sentences and relation descriptions into corresponding embedding. Next, the category-aware calibration module can obtain the unbiased and discriminative prototype representations according to the embeddings of sentences and relation description. We can predict the label of each query sentence based on the similarity between all category prototypes and this query sentence. Finally, the task-aware training strategy module can balance the weight between simple and hard data and then ensure propagation of the correct information.

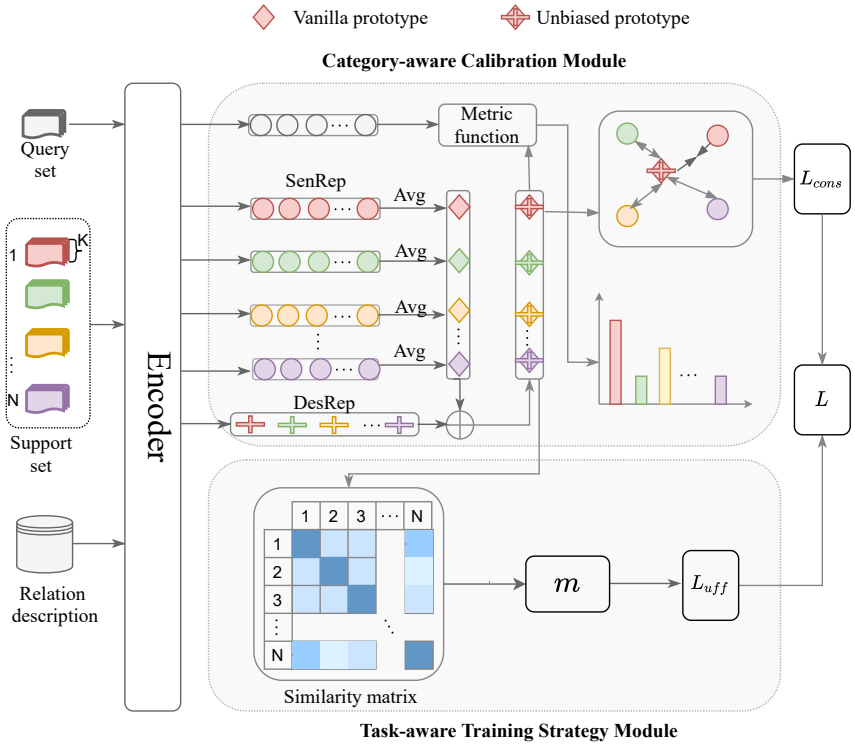


Figure 2. An overview of the few-shot relation extraction framework. Support set follows N-way K-shot setting. Different colors in the support set denote different relations categories. "SenRep" and "DesRep" denote the abbreviations of sentence semantic representation and relation description representation, respectively.

In the next section, we first describe the formulation of the problem in Section 3.1. We then detail the category-aware calibration module in Section 3.2 and the task-aware training strategy module in Section 3.3.

3.1. Task Definition

Relation Extraction. Given an L_x-word text x with a head entity e_h and a tail entity e_t, i.e., $x = \{w_1, \cdots, e_h, \cdots, e_t, \cdots, w_{L_x}\}$, the RE task can be formulated as training a model to predict the relation label r between e_h and e_t, where r belongs to a pre-defined relation label set $\mathcal{R}$. It is worth noting that the entity span may consist of multiple words.

Few-shot Relation Extraction. Few-shot relation extraction aims to identify the emerging novel relation labels without sufficient labeled data. Therefore, the predefined relation label set $\mathcal{R}$ is divided into the base categories R_b and novel categories R_n for the training and test stage, respectively. This setting simulates the test environment, where $R_b \cup R_n = R$ and $R_b \cap R_n = \oslash$. Next, quantities of meta tasks are constructed for few-shot RE. Specifically, a meta task $\mathcal{T}$ consists of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$: $\mathcal{T} = (\mathcal{S}, \mathcal{Q})$. Following the typical N-way K-shot setting of ML learning, the support set $\mathcal{S} = \left\{x_r^j; r = 1, \cdots, N, j = 1, \cdots, K\right\}$ contains N categories, each category with K labeled instances. The query set $\mathcal{Q}$ includes the same N relation categories as $\mathcal{S}$. The few-shot RE methods are trained on meta tasks sampled from the base categories R_b, learn general knowledge, and are tested on other meta tasks sampled from the novel categories R_n.

External Knowledge. The relation description $d_r = \{w_1, \cdots, w_{L_d}\}$ for each relation r is also given, where L_d denotes the word length of d_r.

3.2. Category-Aware Calibration Module

In this section, we first calibrate the representation of each category, and then predict the label of the query sentence by a metric function, which calculates the distance between the query sentence and these categories.

3.2.1. Feature Encoder

We employ E to denote the text feature encoder. We use BERT_{base} as E, as shown in Figure 3. We can then obtain the contextual semantic representation h_x of an instance x:

$$\boldsymbol{h}_x = E(x)[h] \oplus E(x)[t], \tag{1}$$

where $\boldsymbol{h}_x \in \mathbb{R}^{2d}$, d is the embedding dimension of E, $\oplus$ presents the concatenation operation, and $E(x)[h]$ and $E(x)[t]$ denote embedding of the start token of head entity e_h and tail entity e_t, respectively.

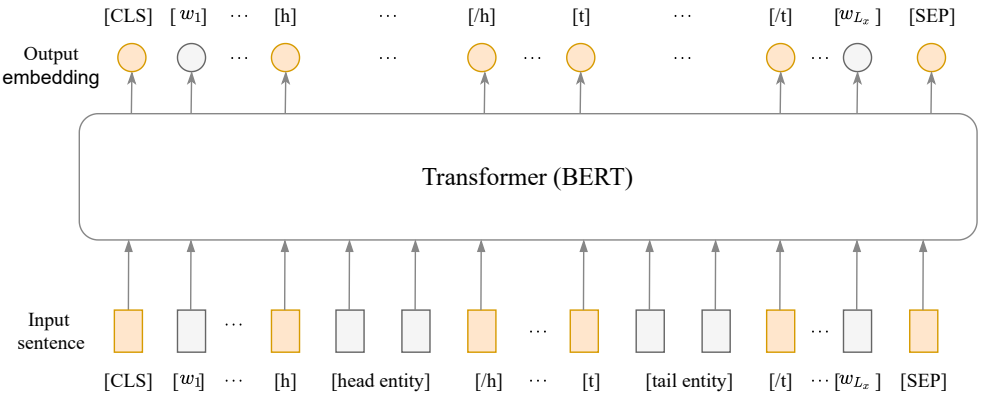


Figure 3. The feature encoder of our proposal. Orange items represent special mark tokens, e.g., [CLS] and [SEP] are the start and end of the input sentence with L_x words, and [h] ([t]) and [/h] ([/t]) are the start and end of the head entity span and tail entity span, respectively.

Additionally, we can gain the relation description representation for each relation r:

$$\boldsymbol{h}_{d_r} = E(d_r)[CLS] \oplus \frac{1}{L_d} \sum_{i=1}^{L_d} E(d_r)[w_i], \tag{2}$$

where $\boldsymbol{h}_{d_r} \in \mathbb{R}^{2d}$, $E(d_r)[CLS] \in \mathbb{R}^d$ demonstrates embedding of the start token of the relation description text, and $E(d_r)[w_i] \in \mathbb{R}^d$ illustrates the embedding of word w_i in the relation description text.

3.2.2. Category Distribution Calibration

Following vanilla prototypical network [7], we average all the instance embeddings in the support set for each relation as vanilla prototype:

$$\boldsymbol{c}_{x_r} = \frac{1}{K}\sum_{j=1}^{K} \boldsymbol{h}_{x_r^j},\ r = 1 \cdots N, \tag{3}$$

where we treat prototype $\boldsymbol{c}_{x_r}$ as the representation of category r. However, $\boldsymbol{c}_{x_r}$ is vulnerable to outliers in the few-shot scenario where there are only very few instances for demonstration, leading to semantic distribution and discrimination bias.

Fortunately, the relation description summarizes the semantic characteristics, which elaborates the real meaning. Therefore, we leverage the relation description to calibrate the distribution of the corresponding category representation:

$$\boldsymbol{c}_r = \boldsymbol{c}_{x_r} + \boldsymbol{h}_{d_r}, \tag{4}$$

compared to $\boldsymbol{c}_{x_r}$, $\boldsymbol{c}_r$ is an unbiased prototype and much closer to the real distribution of relation r. In other words, we calibrate the category distribution without introducing supernumerary parameters. We can then predict the category of a query instance by the following metric function:

$$p(y = r|q) = -\frac{\exp\left(-d\left(\boldsymbol{c}_r, \boldsymbol{h}_q\right)\right)}{\sum_{r'=1}^{N} \exp\left(-d\left(\boldsymbol{c}_{r'}, \boldsymbol{h}_q\right)\right)}, \tag{5}$$

where $p(y = r|q)$ means the probability of query q belonging to relation r. $\boldsymbol{h}_q$, yielded from Equation (1), is a representation of the query sentence. $d(\cdot, \cdot)$ is the distance function based on dot production. Subsequently, we apply cross-entropy loss to optimize prototype representation:

$$L_{ce} = -\sum_{q \in Q} \mathbb{I}_r \log(p(y = r|q)), \tag{6}$$

where $\mathbb{I}_r$ is an indication function, $\mathbb{I}_r = 1$ when query q belongs to relation r, otherwise $\mathbb{I}_r = 0$.

3.2.3. Category Discrimination Calibration

As for semantic discrimination bias, we apply contrastive learning to discriminate the representations of instances for each relation. In detail, instances should be close to the prototype belonging to the same category and far away from other prototypes as follows:

$$L_{\text{cons}} = \frac{-1}{NK}\sum_{r=1}^{N} log\frac{\sum_{j=1}^{K} exp(\mathbf{c}_r \cdot \mathbf{h}_{x_r^j}/\tau)}{\sum\limits_{r' \neq r}\sum_{j=1}^{K}(\mathbf{c}_{x_{r'}} \cdot \mathbf{h}_{x_r^j}/\tau)}, \tag{7}$$

where τ is a temperature hyper-parameter, $\mathbf{h}_{x_r^j}$ denotes the representation of j-th instance in relation r. Thus, here we can obtain discriminative prototypes based on Equation (7).

3.3. Task-Aware Training Strategy Module

For a sentence with true relation label r, we predict that it belongs to r with a confidence of $p(y = r|q)$ by Equation (5). We define the easily classified sentence as a very simple instance when $p(y = r|q) \to 1$. Conversely, the extremely difficult classified sentence is the

very hard instance with very low confidence. The task-aware training strategy module is designed to optimize our proposal with moderately hard data.

3.3.1. Hard Instances

Intuitively, the models will benefit if they focus more on hard instances instead of treating all instances equally [13]. Therefore, we apply the focal loss function [13] on hard instance to modify the cross entropy loss L_{focal}:

$$L_{focal} = -\sum_{q \in Q} \mathbb{I}_r \log\big((1 - p(y = r|q)^{\varphi})\, p(y = r|q)\big), \tag{8}$$

where $\varphi > 0$ is a hyper-parameter [13] and reduces the relative loss contributed by very simple instances. Furthermore, Equation (8) is cross-entropy loss when $\varphi = 0$.

3.3.2. Hard Meta Tasks

However, the harder the sentence, the higher the L_{focal} weight assigned to this sentence, which may lead to TAPN failing to learn knowledge since L_{focal} focuses excessively on very hard sentences. Therefore, we design an inverse focal loss function at the meta-task level, which pays less attention to the very hard task consisting of very hard classified sentences. We can observe that the greater the inter-class similarity in a meta task, the harder this meta task becomes. We then use the inter-class similarity matrix $\boldsymbol{M} \in \mathbb{R}^{N \times N}$ to measure the difficulty (hardness) of meta task:

$$\mathbf{m}_{i,j} = \frac{\mathbf{c}_i \cdot \mathbf{c}_j}{\|\mathbf{c}_i\| \times \|\mathbf{c}_j\|}, \tag{9}$$

where $\|\cdot\|$ represents the Euclidean norm. Next, we use a scalar m to determine the hard magnitude of a specific meta task in the current mini-batch as follows:

$$m_b = \frac{\exp(\|\boldsymbol{M}_b\|_F)}{\sum_{b'=1}^{B} \exp\big(\|\boldsymbol{M}_{b'}\|_F\big)}, \tag{10}$$

where B is the batch size in training stage, m_b is the difficulty of b-th meta task in current batch, and $\|\cdot\|_F$ is the Frobenius norm. The task-aware loss is then defined as follows:

$$L_{uff} = -\sum_{b=1}^{B} \frac{1}{m_b} \sum_{q \in Q} \mathbb{I}_r \log\big((1 - p(y = r|q)^{\varphi})\, p(y = r|q)\big), \tag{11}$$

L_{uff} pays less attention to a hard meta task but focuses on hard instances in the meta task. Namely, L_{uff} balances the weight between easy and hard data; therefore, it can learn useful knowledge from moderately hard data.

Lastly, the final objective loss is designed as follows:

$$L = L_{cons} + L_{uff} \tag{12}$$

4. Experiments

In this section, we first discuss several research questions in Section 4.1. We then introduce the dataset and baselines to compare with those in Sections 4.2 and 4.3, respectively. Finally, we provide some implementation details in Section 4.4.

4.1. Research Questions

We design the following research questions to guide our experiments and examine the effectiveness of our proposal.

- **RQ1**: Does our proposal outperform the state-of-the-art baselines in terms of accuracy for few-shot relation extraction?

- **RQ2**: How does sentence length influence the performance of our proposal and baselines?
- **RQ3**: What is the impact of the different components of TAPN?
- **RQ4**: Are the results interpretable from the view of error analysis?

4.2. Datasets

We conduct our experiment on FewRel [4]. There are 64, 20 and 16 relations for training, validation and testing, respectively. Since the 20 test relations are not reported, we re-split the original published relations into 50, 14 and 16 for training, validation and testing, respectively, according to existing methods [10,11]. In addition, the statistics of FewRel are listed in Table 1. Moreover, the test relation descriptions are listed in Table 2.

Table 1. Statistics of FewRel. "#Rel" and "#Instance" denote the number of relations and instances, respectively. "Length" means the average token length of instances.

Task	#Rel	#Instance	Length
Training	50	35,000	25
Validation	14	9800	24
Testing	16	11,200	24

Table 2. Relation descriptions for the test data.

Id	Relation Name	Relation Description
"follows"	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 is 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.

4.3. Model Summary

We introduce two group competitive baselines for the few-shot RE task to be compared with. We first illustrate the basic ML methods:

- **Snail** [41] applies the temporal convolutions to aggregate information from past experience and designs a soft attention mechanism to pinpoint specific pieces of information.

- **GNN** [42] defines a graph neural network architecture to propagate label information from labeled data to unlabeled data.
- **Siamese** [43] uses two twin networks with shared weights to calculate the similarity of two inputs and then determines whether they belong to the same category.
- **Proto** [7] predicts the relation labels by calculating the similarity between the query sentences and the prototype of each category, which is obtained by averaging the representations of all instances belonging to a specific category.
- **BERT-PAIR** [44] concatenates the query instance with all supporting instances with a particular label as a series of sequences, and then calculates the similarity of two pairs of instances for predicting the relation of query instance.

We then list some improved prototypical networks by introducing external knowledge through carefully designed complex modules:

- **KEFDA** [45] designs a knowledge-enhanced prototypical network to conduct instance matching and a relation-meta learning network for implicit relation matching.
- **ConceptFERE** (We use the ConceptFERE(simple) version here to allow for computation overheads. Ref. [11] develops a self-attention-based fusion module to incorporate sentence embedding and entity concept embedding, which is valuable for the relation classifier.
- **HCRP** [12] introduces external relation description and combines global and local features as hybrid prototypes, which better learn representations by exploiting relation label information.

Finally, we present the model proposed in this paper:

- **TAPN** leverages the relation description to calibrate the prototype representation without introducing extra parameters and designs an effective training strategy to optimize the model.

4.4. Implementation Details

The model configurations are kept the same across all models discussed, including our proposal and the selected baselines. In detail, following [4,46], we assess the performance of DRK on four classic meta-tasks: 5-way 1-shot, 5-way 5-shot, 10-way 1-shot and 10-way 5-shot. We apply BERT_{base} as the feature encoder and use ADAM to optimize all the models. In addition, we follow the parameter setting of FewRel [4] and tune other hyperparameters through performing a grid search on a validation set. Furthermore, we present the parameter settings in Table 3. It is worth noting that we set τ to 0.4 on the 10-way 1-shot meta task through conducting a grid search on a validation set.

Table 3. Hyperparameters of our proposal.

Hyperparameters	Set
Training iterations	20,000
Validation iterations	2000
Test iterations	10,000
learning rate	2×10^{-5}
sentence length L_x	128
d	768
τ	1
φ	1
batch size	4
grad iter	1

5. Results and Discussion

5.1. Overall Evaluation

For answering **RQ1**, we assess the RE performance of TAPN along with eight competing baselines on four meta-tasks. The overall results in terms of accuracy are listed in Table 4.

Generally, for meta tasks with the same shot number, the performance of all models deteriorates as the number of relational categories (ways) increases. In addition, for the same way-number meta tasks, all the models achieve better performance as the shot-number increases. The above phenomenon indicates that the difficulty of the relation-extraction task increases as the number of shots reduces and the number of ways increases. This can contribute to under-fitting for test tasks, which suffers from a lack of data.

Table 4. Overall performance of our proposal and baselines in terms of average accuracy(%) on four typical meta-tasks. The results of the best baseline and the best performer in each column are underlined and boldfaced, respectively. Statistical significance of pairwise differences of the best baseline against our proposed TAPN is determined via a *t*-test (▲ for $p < 0.05$). † marks the results quoted from the original published papers.

Model	Avg	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
Snail	64.22	57.82	80.53	50.40	68.11
GNN	67.62	66.48	82.65	48.14	73.22
Siamese	80.40	81.29	88.18	71.00	81.12
Proto	78.36	78.59	88.99	64.07	81.80
BERT-PAIR	81.69	82.57	89.00	73.37	81.81
KEFDA	80.02	80.46	89.88	68.23	81.49
ConceptFERE †	82.61	84.28	90.34	74.00	81.82
HCRP	84.14	<u>86.82</u>	<u>90.38</u>	<u>75.98</u>	<u>83.39</u>
TAPN(ours)	87.44	**90.98** ▲	**91.76** ▲	**80.66** ▲	**86.37** ▲

Subsequently, we focus on the baseline. For the first group methods, BERT-PAIR achieves the best results due to the carefully designed model structure. For the second group baselines with external knowledge, most models achieve better performance than first group models. In addition, HCRP is the best baseline on four meta tasks. This demonstrates that external knowledge provides rich information to alleviate the few-shot dilemma.

Next, we focus on the performance of our proposal on four meta tasks. Generally, our suggested TAPN is superior to all baselines on all meta-tasks and gains a 3.30% improvement in average accuracy, which confirms the validity of the TAPN. In detail, TAPN exhibits 1.38%, 4.16%, 2.98% and 4.68% improvements in the accuracy of HCRP on 5-way 5-shot, 5-way 1-shot, 10-way 5-shot and 10-way 1-shot meta-tasks, respectively, and the performance growth of our proposal increases as the way-number grows and the shot-number reduces. This demonstrates that TAPN can capture unbiased and discriminative features in the harsh few-shot scenario. In addition, we also evaluate the performance precision of our proposed method and the state-of-the-art baseline HCRP in Table 5. We can observe that our proposed method still outperforms HCRP by 2% improvement in terms of average precision.

Table 5. The performance of our proposal and HCRP in terms of precision(%) on test relations.

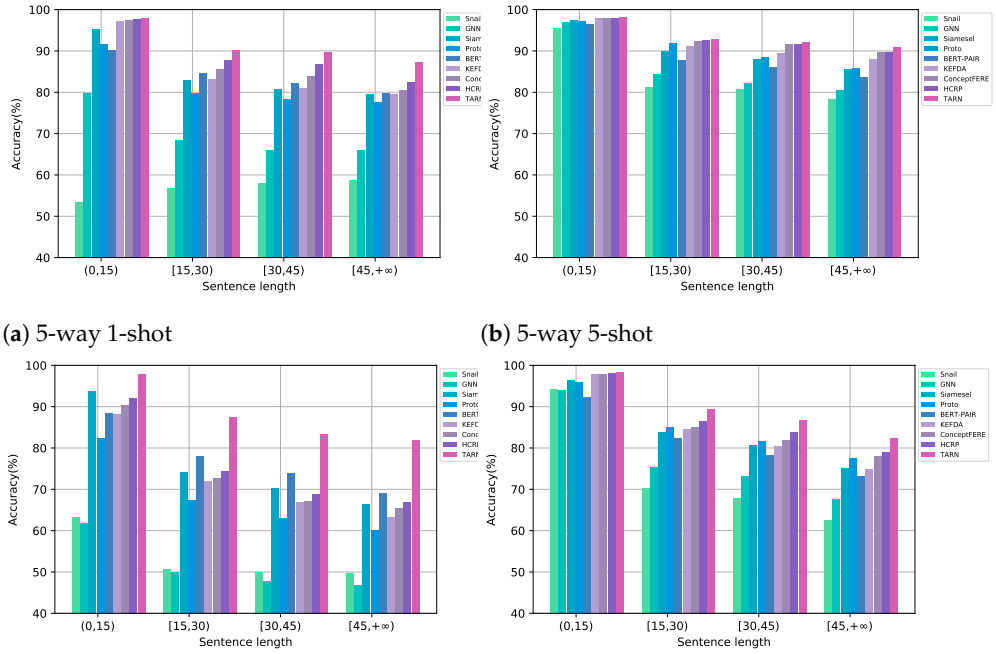
Relation	TAPN	HCRP
average	0.83	0.81
main subject	0.85	0.94
sport	0.76	0.59
constellation	0.79	0.69
member of	0.75	0.89
position played on team	0.83	0.91
voice type	0.98	0.79
military rank	0.85	0.78
child	0.72	0.96
original language of film or TV show	0.79	0.86
part of	0.93	0.93
spouse	0.83	0.64
mother	0.61	0.81
competition class	0.96	0.91
located in or next to body of water	0.89	0.95
crosses	0.97	0.97
follows	0.80	0.42

5.2. Sentence Length

As for **RQ2**, we study the influence of sentence length on the behavior of all models, in accordance with the sentence length L_s. In detail, considering the distribution of testing data, we group the sentences into four groups, i.e., $L_s \in (0, 15), [15, 30), [30, 45), [45, +\infty)$. The results are plotted in Figure 4.

Generally, almost all the performances of the models drop with an increase in sentence length, which can be clearly observed in Figure 4c. This phenomenon may be caused by the model failing to capture key information as the sentence length increases. In addition, long sentences are more likely to introduce noise.

Next, we compare the results of our proposed method against the baselines. Furthermore, we take the worst-performing 10-way 1-shot meta task for instance to analyze the results. We find that our proposal obtains the best results at every sentence length on all four meta tasks. Furthermore, our proposal is less sensitive to the length of the input sentence than other baselines. For instance, compared to the best baseline HCRP degrades by 24.98% from 91.97% at $L_s \in (0, 15)$ to 66.99% at $L_s \in [45, +\infty)$, our proposed TAPN only decreases by 15.93% from 97.36% at $L_s \in (0, 15)$ to 82.01% at $L_s \in [45, +\infty)$. In addition, the improvement magnitude of TAPN consistently increases along with an increasing length of the input sentence, e.g., TAPN outperforms the best baseline HCRP by an improvement of 5.96%, 13.25%, 14.64% and 15.02% at $L_s \in (0, 15), [15, 30), [30, 45), [45, +\infty)$, respectively. This demonstrates that our proposal can capture discriminative features to alleviate the noise caused by long and tedious sentences. Similar results can be observed for the 10-way 5-shot, 5-way 5-shot, and 5-way 1-shot meta tasks.



(**a**) 5-way 1-shot (**b**) 5-way 5-shot

(**c**) 10-way 1-shot (**d**) 10-way 5-shot

Figure 4. Effect on the performance of our proposed method and baselines affected by sentence length on four typical meta tasks: 5-way 1-shot, 5-way 5-shot, 10-way 1-shot and 10-way 5-shot.

5.3. Ablation Study

For **RQ3**, we perform an ablation study to understand the contribution of the various components of our proposal. In the ablation study, we replace or remove some specific components to measure their influence on TAPN, which is marked with the notation"wo". Specifically, "wo/rel" and "wo/cons" denote removal of the category distribution calibration in Section 3.2.2 and category dicriminative calibration in Section 3.2.3, respectively. The "wo/task" and "wo/instance" refer to removal of the hard meta-task finding component in Section 3.3.2 and hard instance-finding component in Section 3.3.1, respectively. It is worth noting that we only conduct an ablation study on 5-way 1-shot and 5-way 5-shot meta tasks given the high computation cost on 10-way 1-shot and 10-way 5-shot meta tasks. Furthermore, the results are presented in Table 6.

As displayed in Table 6, the removal of components leads to model degeneration, proving the efficacy of each component. Additionally, "wo/rel" leads to the biggest drop among the four components as marked in Table 6. The "wo/rel" plays the most important role, which verifies that the previous prototype representation is vulnerable in the few-shot scenario, which affects subsequent classification accuracy. Furthermore, the category distribution calibration module calibrates prototype representation to be unbiased and discriminative without introducing extra parameters.

Table 6. Ablation study of TAPN for the 5-way 1-shot and 5-way 5-shot meta tasks.

Model	5-Way 1-Shot	5-Way 5-Shot
TAPN	90.98	91.76
wo/rel	83.38↓	89.89↓
wo/cons	87.91	92.17
wo/task	89.02	91.11
wo/instance	85.02	90.95

5.4. Error Analysis

To answer **RQ4**, we first analyze the accuracy of each test relation, and then determine the error sources via error analysis.

First, we present the accuracy of the best baseline HCRP and our proposal on each test relation in Figure 5. Specifically, following Brody et al. [47], we use the parameters of 10-way 5-shot to evaluate the performance on test data by relation. Specifically, for each test relation, we randomly select 5 examples (that is, K = 5) and 50 examples of that relation and place them into the support set and query set, respectively. As displayed in Figure 5, we can observe that the performance of our model is more stable than HCRP. The good performance of HCRP is contributed to by some easily distinguished relations but fails on some difficulty relations, e.g., the accuracy of HCRP is under 40% on relation P206, P26 and P641. Fortunately, our proposed method performs well across all relation categories, and the accuracy on every relation is over 40%.

Next, we conduct an error analysis on the relation "follows" to determine the error sources and the findings are summarized in Table 7. Generally, TAPN outperforms HCRP by 4% improvement in accuracy on the relation "follows". On the one hand, TAPN reduces the error source, e.g., relation "constellation". This may contribute to the calibration based on the relation description shifting the prototype away from an irrelevant relation category. On the other hand, TAPN decreases the error probability on the relation "part of", meaning that TAPN can capture the discriminative features of each relation.

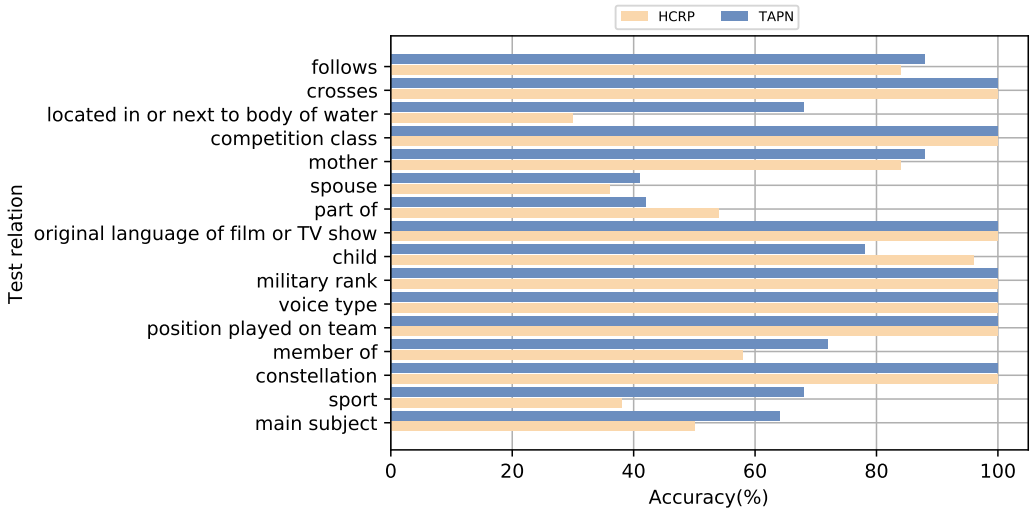


Figure 5. The accuracy of our proposed TAPN method and the best baseline HCRP on each test relation.

Table 7. The error source of relation "follows".

Model	"follows"	Other Relations
HCRP	0.88	"part of": 0.14, "constellation": 0.02
TAPN	0.84	"part of": 0.12

6. Conclusions and Future Work

In this paper, we propose a taxonomy-aware prototypical network to solve the few-shot relation extraction. Specifically, we design a category-aware calibration module that utilizes the relation description and contrastive learning to calibrate prototype representation to become sufficiently unbiased and discriminative. Furthermore, we develop a task-aware training strategy module, which dynamically balances the weight of easy and hard tasks. In addition, we conduct extensive experimentation on FewRel for four typical meta tasks.

The results demonstrate that our proposal exceeds the state-of-the-art baseline in average accuracy.

However, our proposal may be limited to addressing the cross-domain relation extraction task, where the testing and training data originate from various domains. Therefore, regarding the feature work, on the one hand, we plan to examine the generalization of TAPN in the cross-domain few-shot scenario [48]. On the other hand, we would like to introduce prompt learning for the true few-shot [49] scenario, where both training and validation data are scarce. For example, we can design a template to close the gap between relation extraction and the pre-trained language model, which can exploit common knowledge learned from pre-trained language models.

Author Contributions: Write the original manuscript, M.W.; review and edit manuscript, J.Z. and H.C. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the first author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nayak, T.; Majumder, N.; Goyal, P.; Poria, S. Deep Neural Approaches to Relation Triplets Extraction: A Comprehensive Survey. *Cogn. Comput.* **2021**, *13*, 1215–1232. [CrossRef]
2. Bassignana, E.; Plank, B. What Do You Mean by Relation Extraction? A Survey on Datasets and Study on Scientific Relation Classification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Dublin, Ireland, 22–27 May 2022; pp. 67–83.
3. Wang, H.; Qin, K.; Zakari, R.Y.; Lu, G.; Yin, J. Deep neural network-based relation extraction: An overview. *Neural Comput. Appl.* **2022**, *34*, 4781–4801. [CrossRef]
4. Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; Sun, M. 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, Brussels, Belgium, 31 October–4 November 2018; pp. 4803–4809.
5. Wen, W.; Liu, Y.; Ouyang, C.; Lin, Q.; Chung, T.L. Enhanced prototypical network for few-shot relation extraction. *Inf. Process. Manag.* **2021**, *58*, 102596. [CrossRef]
6. Huang, W.; He, M.; Wang, Y. A Survey on Meta-learning Based Few-Shot Classification. In Proceedings of the Machine Learning and Intelligent Communications-6th EAI International Conference, MLICOM 2021, Wuzhou, China, 17–18 November 2021; Volume 438; pp. 243–253.
7. Snell, J.; Swersky, K.; Zemel, R.S. Prototypical Networks for Few-shot Learning. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 4077–4087.
8. Gao, T.; Han, X.; Liu, Z.; Sun, M. Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification. In Proceedings of the The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Hilton Hawaiian Village, HI, USA, 27 January–1 February 2019; pp. 6407–6414.
9. Qu, M.; Gao, T.; Xhonneux, L.A.C.; Tang, J. Few-shot Relation Extraction via Bayesian Meta-learning on Relation Graphs. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event, 13–18 July 2020; PMLR, 2020; Volume 119, pp. 7867–7876.
10. Yang, K.; Zheng, N.; Dai, X.; He, L.; Huang, S.; Chen, J. Enhance Prototypical Network with Text Descriptions for Few-shot Relation Classification. In Proceedings of the CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, 19–23 October 2020; pp. 2273–2276.
11. Yang, S.; Zhang, Y.; Niu, G.; Zhao, Q.; Pu, S. 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, ACL/IJCNLP 2021, Virtual Event, 1–6 August 2021; pp. 987–991.
12. Han, J.; Cheng, B.; Lu, W. Exploring Task Difficulty for Few-Shot Relation Extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021; Moens, M., Huang, X., Specia, L., Yih, S.W., Eds.; 2021; pp. 2605–2616.

13. Lin, T.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
14. Shang, Y.; Huang, H.; Sun, X.; Wei, W.; Mao, X. A pattern-aware self-attention network for distant supervised relation extraction. *Inf. Sci.* **2022**, *584*, 269–279. [CrossRef]
15. Xu, J.; Chen, Y.; Qin, Y.; Huang, R.; Zheng, Q. A Feature Combination-Based Graph Convolutional Neural Network Model for Relation Extraction. *Symmetry* **2021**, *13*, 1458. [CrossRef]
16. Bhamare, B.R.; Prabhu, J. A supervised scheme for aspect extraction in sentiment analysis using the hybrid feature set of word dependency relations and lemmas. *PeerJ Comput. Sci.* **2021**, *7*, e347. [CrossRef] [PubMed]
17. Ravi, S.; Larochelle, H. Optimization as a Model for Few-Shot Learning. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
18. Yu, Y.; Wang, G.; Ren, H.; Cai, Y. Incorporating Bidirection-Interactive Information and Semantic Features for Relational Facts Extraction (Student Abstract). In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; pp. 15947–15948.
19. Ghosh, D.; Muresan, S. Relation Classification using Entity Sequence Kernels. In Proceedings of the COLING 2012, Mumbai, India, 8–15 December 2012; pp. 391–400.
20. Leeuwenberg, A.; Buzmakov, A.; Toussaint, Y.; Napoli, A. Exploring Pattern Structures of Syntactic Trees for Relation Extraction. In Proceedings of the Formal Concept Analysis—13th International Conference, Neja, Spain, 23–26 June 2015; Volume 9113, pp. 153–168.
21. Cho, C.; Choi, Y.S. Dependency tree positional encoding method for relation extraction. In Proceedings of the SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, 22–26 March 2021; pp. 1012–1020.
22. Shi, Y.; Xiao, Y.; Quan, P.; Lei, M.; Niu, L. Distant Supervision Relation Extraction via adaptive dependency-path and additional knowledge graph supervision. *Neural Netw.* **2021**, *134*, 42–53. [CrossRef] [PubMed]
23. Reichartz, F.; Korte, H.; Paass, G. Composite Kernels For Relation Extraction. In Proceedings of the ACL 2009, 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–7 August 2009; pp. 365–368.
24. Wen, H.; Zhu, X.; Zhang, L.; Li, F. A gated piecewise CNN with entity-aware enhancement for distantly supervised relation extraction. *Inf. Process. Manag.* **2020**, *57*, 102373. [CrossRef]
25. Niu, W.; Chen, Q.; Zhang, W.; Ma, J.; Hu, Z. GCN2-NAA: Two-stage Graph Convolutional Networks with Node-Aware Attention for Joint Entity and Relation Extraction. In Proceedings of the ICMLC 2021: 13th International Conference on Machine Learning and Computing, Shenzhen, China, 26 February–1 March 2021; pp. 542–549.
26. Geng, Z.; Chen, G.; Han, Y.; Lu, G.; Li, F. Semantic relation extraction using sequential and tree-structured LSTM with attention. *Inf. Sci.* **2020**, *509*, 183–192. [CrossRef]
27. Peng, Y.; Rios, A.; Kavuluru, R.; Lu, Z. Chemical-protein relation extraction with ensembles of SVM, CNN, and RNN models. Proceedings of the BioCreative VI Workshop *arXiv* **2018**, arXiv:1802.01255.
28. Lee, H.; Li, S.; Vu, T. Meta Learning for Natural Language Processing: A Survey. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 666–684.
29. Li, Y.; Ma, Z.; Gao, L.; Wu, Y.; Xie, F.; Ren, X. Enhance prototypical networks with hybrid attention and confusing loss function for few-shot relation classification. *Neurocomputing* **2022**, *493*, 362–372. [CrossRef]
30. Obamuyide, A.; Vlachos, A. Model-Agnostic Meta-Learning for Relation Classification with Limited Supervision. In Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5873–5879.
31. Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; Lillicrap, T.P. Meta-Learning with Memory-Augmented Neural Networks. In Proceedings of the 33nd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; Volume 48, pp. 1842–1850.
32. Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1126–1135.
33. Song, Y.; Liu, Z.; Bi, W.; Yan, R.; Zhang, M. Learning to Customize Model Structures for Few-shot Dialogue Generation Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 5–10 July 2020; pp. 5832–5841.
34. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.S.; Hospedales, T.M. Learning to Compare: Relation Network for Few-Shot Learning. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1199–1208.
35. Zheng, H.; Zhou, Y.; Huang, X. Improving Cancer Metastasis Detection via Effective Contrastive Learning. *Mathematics* **2022**, *10*, 2404. [CrossRef]
36. Wu, Z.; Wang, S.; Gu, J.; Khabsa, M.; Sun, F.; Ma, H. CLEAR: Contrastive Learning for Sentence Representation. *CoRR* **2020**. [CrossRef]
37. Meng, Y.; Xiong, C.; Bajaj, P.; Tiwary, S.; Bennett, P.; Han, J.; Song, X. COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining. *CoRR* **2021**. [CrossRef]

38. Yan, Y.; Li, R.; Wang, S.; Zhang, F.; Wu, W.; Xu, W. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, 1–6 August 2021; pp. 5065–5075. [CrossRef]
39. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6894–6910. [CrossRef]
40. Jiang, T.; Huang, S.; Zhang, Z.; Wang, D.; Zhuang, F.; Wei, F.; Huang, H.; Zhang, L.; Zhang, Q. PromptBERT: Improving BERT Sentence Embeddings with Prompts. *CoRR* **2022**. [CrossRef]
41. Mishra, N.; Rohaninejad, M.; Chen, X.; Abbeel, P. A Simple Neural Attentive Meta-Learner. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
42. Satorras, V.G.; Estrach, J.B. Few-Shot Learning with Graph Neural Networks. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
43. Müller, T.; Pérez-Torró, G.; Franco-Salvador, M. Few-Shot Learning with Siamese Networks and Label Tuning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL, Dublin, Ireland, 22–27 May 2022; pp. 8532–8545.
44. Gao, T.; Han, X.; Zhu, H.; Liu, Z.; Li, P.; Sun, M.; Zhou, J. 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, Hong Kong, China, 3–7 November 2019.
45. Zhang, J.; Zhu, J.; Yang, Y.; Shi, W.; Zhang, C.; Wang, H. Knowledge-Enhanced Domain Adaptation in Few-Shot Relation Classification. In Proceedings of the KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, 14–18 August 2021; pp. 2183–2191.
46. Han, Y.; Qiao, L.; Zheng, J.; Kan, Z.; Feng, L.; Gao, Y.; Tang, Y.; Zhai, Q.; Li, D.; Liao, X. Multi-view Interaction Learning for Few-Shot Relation Classification. In Proceedings of the 30th ACM International Conference on Information and Knowledge Management, Gold Coast, Australia, 1–5 November 2021.
47. Brody, S.; Wu, S.; Benton, A. Towards Realistic Few-Shot Relation Extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 5338–5345.
48. Tseng, H.; Lee, H.; Huang, J.; Yang, M. Cross-Domain Few-Shot Classification via Learned Feature-Wise Transformation. In Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
49. Perez, E.; Kiela, D.; Cho, K. True Few-Shot Learning with Language Models. In Proceedings of the Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, Online, 6–14 December 2021; pp. 11054–11070.