Domain-aware and Co-adaptive Feature Transformation for Domain Adaption Few-shot Relation Extraction

Yijun Liu [1,2,3], **Feifei Dai**[1,3(✉)], **Xiaoyan Gu**[1,2,3(✉)], **Minghui Zhai**[1,2,3]
Bo Li[1,3], **Meiou Zhang**[4]
[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] Key Lab of Cyberspace Security Defense, Beijing, China
[4] China Mobile (Shenzhen) Co., Ltd. Shenzhen, China
{liuyijun, daifeifei, guxiaoyan, zhaiminghui, libo}@iie.ac.cn
zhangmeiouit@chinamobile.com

Abstract

Few-shot relation extraction (FSRE) can alleviate the data scarcity problem in relation extraction. However, FSRE models often suffer a significant decline in performance when adapting to new domains. To overcome this issue, many researchers have focused on domain adaption FSRE (DAFSRE). Nevertheless, existing approaches primarily concentrate on the source domain, which makes it difficult to accurately transfer useful knowledge to the target domain. Additionally, the lack of distinction between relations further restricts the model performance. In this paper, we propose the domain-aware and co-adaptive feature transformation approach to address these issues. Specifically, we introduce a domain-aware transformation module that leverages the target domain distribution features to guide the domain-aware feature transformations. This can enhance the model's adaptability across domains, leading to improved target domain performance. Furthermore, we design co-adaptive prototypical networks to perform co-adaptive feature transformation through a transformer mechanism. This results in more robust and distinguishable relation prototypes. Experiments on DAFSRE benchmark datasets demonstrate the effectiveness of our method, which outperforms existing models and achieves state-of-the-art performance.

Keywords: relation extraction, few-shot learning, domain adaption

1. Introduction

Relation extraction (RE) is a fundamental task in information extraction, aiming to identify semantic relations between entities within a given text. It has widespread applications in various fields (Li et al., 2022b; Zhu et al., 2023; Yao and Lai, 2023). However, RE models often suffer from data scarcity problems due to the expensive labor required for manual annotation. To alleviate this problem, many researchers (Han et al., 2018; Gao et al., 2019) (Soares et al., 2019) turn to the task of few-shot RE (FSRE). FSRE involves classifying a query instance by exploring only a few labeled examples. Despite the remarkable progress, when adapting to the new domain, the FSRE models' performance would decline significantly.

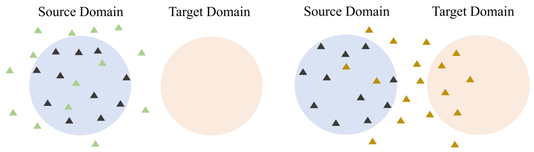
To solve the above problems, Gao et al. (Gao et al., 2019) first introduced the task of domain adaptation FSRE (DAFSRE). Compared to FSRE, DAFSRE aims to effectively transfer knowledge from the source domain to the target domain by exploring a few labeled instances. Thus, the model can use this knowledge to guide the RE in the target domain. As shown in Table 1, in the source domain, the model needs to learn the measure of similarity scores between each relation ("has part" and "instance of") and query instance, then perform classification based on these similarity scores. In the target domain, this model should identify which relation ("inheritance type of" or "causative agent of") the query instance belongs to, with the help of the learned measure ability in the source domain.

Current approaches for DAFSRE primarily rely on meta-learning, which aims to train the model to learn how to learn (Vinyals et al., 2016). It achieves this by sampling few-shot classification tasks from the source domain and optimizing the model to perform well on the target domain. Therefore, the model can acquire cross-domain knowledge and leverage it to quickly adapt to the target domain (Soares et al., 2019; Zhai et al., 2023; Liu et al., 2023). However, these methods cannot effectively develop the potentially exploitable knowledge, such as the knowledge from the target domain which may provide valuable insights for domain adaptation. To this end, several scholars propose the adversarial training-based methods (Wang et al., 2022; Li et al., 2022a; Chen et al., 2023b), which suggests leveraging unlabeled target domain data and conducting adversarial training during meta-learning. As these methods have been found to be unstable sometimes (Sajeeda and Hossain, 2022) , several studies (Zhang et al., 2022; Hu and Ma, 2022) propose the feature transformation-based approaches to use feature-wise transformation layers for do-

LREC-COLING 2024, pages 5275–5285
20-25 May, 2024.

Table 1: An example of DAFSRE. The head and tail entity are indicated by blue and red, respectively.

Source Domain (general domain)		
Support Set	(R1) has part	It is native to the Americas, including Central and South America, and ...
	(R2) instance of	14 singles claimed the top spot, including "Poker Face", which started ...
Query Set	(R1) or (R2)?	... accolades including Filmfare Awards nominations for Best Director and Best Screenplay.
Target Domain (medical domain)		
Support Set	(R1) inheritance type of	Tar syndrome is inherited in an autosomal recessive manner and results ...
	(R2) causative agent of	Rickettsioses are caused by obligate ... within the genus rickettsia, mainly ...
Query Set	(R1) or (R2)?	Malignant hyperthermia (mh) is an autosomal dominant metabolic myopathy.



(a) Domain-agnostic Feature Transformation (b) Domain-aware Feature Transformation

Figure 1: The black, green, and brown triangles correspond to source domain features, features enhanced by domain-agnostic feature transformation and by domain-aware feature transformation. The domain-agnostic feature transformation can not broaden the feature to the target domain, which makes less transferable knowledge being learned.

main adaption. The purpose of this method is to enhance the source domain features by applying affine transforms, which can simulate broader variations of feature distributions during the training stage. These approaches can improve the cross-domain learning ability of the model significantly. More recently, the large language models (LLMs) demonstrated great potential in few-shot learning tasks (Brown et al., 2020), which achieve the best results among numerous natural language processing (NLP) benchmarks.

Although these methods have achieved a lot of progress, there are still some unresolved issues that require attention.

First, most existing methods implicitly instruct the broadening of the feature distribution through adversarial training (Tseng et al., 2019; Hu and Ma, 2022) or statistics information (Zhang et al., 2022) to alleviate the domain adaption problem. However, since these domain-agnostic methods primarily focus on the source domain, and sometimes the domain gap can be quite significant, they may not be able to accurately transfer useful knowledge from the source domain to the target domain. Consequently, the model's cross-domain generalization ability cannot be effectively improved. As shown in Figure 1(a), these methods can only broaden the features around the source domain and struggle to align them with the target domain features , which makes less useful and transferable knowledge being learned. As for LLMs, despite their substantial knowledge acquired from extensive corpora, when adapting to some specialized domains, they may experience severe model hallucinations (Zhang et al., 2023b,a) due to their limited understanding of the domain-specific text. Therefore, the model may perform poorly in the target domain.

Second, the current approaches usually concentrate solely on capturing the similarity information between the query and support instances, then, directly perform classification based on this information. This may lead to misclassification when there are confusing relations in the support set. For instance, consider the source domain data presented in Table 1. A query instance containing the keyword "including" may correspond to either the relation classes "has part" or "instance of". In this scenario, a model that exclusively focuses on capturing the similarity information between the query and support instances could potentially generate a high similarity score for the query and both "has part" and "instance of" relations. Consequently, without learning the difference between the "has part" and "instance of" relations, it could be challenging to determine the specific relation class of the query instance.

To tackle these problems, this paper proposes the Domain-aware and Co-adaptive Feature Transformation (DCFT) approach. DCFT aims to mitigate the domain gap and facilitate learning for distinguishing confusing relations. Figure 2 shows the overall framework of DCFT. Specifically, DCFT comprises three components: the encoder, the Domain-aware Transformation Module (DTM), and the Co-adaptive Prototypical Networks (CPN). 1) The encoder is used to convert each word in a sentence into a contextualized embedding. 2) The DTM utilizes the distribution feature of the target set in the unsupervised target domain data to guide a domain-aware feature transformation layer. This layer enhances the source domain data to make it align more closely with the target domain. By doing so, the model is effectively trained on the data related to target domain features. Hence, more useful and transferable knowledge can be learned, resulting in better model performance in the target domain. 3) To emphasize the differentiation among support relation classes, the CPN employs the Transformer mechanism (Vaswani et al., 2017) to model interactions among all samples in the support and query sets. The intra-class aggregation layer and inter-class adaptive layer in these networks can help the model select informative instances and highlight the key feature that aids in differentiating various relation prototypes. Therefore, the model can effectively distinguish different classes.

In summary, the main contributions of this paper are as follows:

- To accurately transfer knowledge from the source to the target domain, we design a domain-aware transformation module to leverage the target domain distribution information, enhancing the model's domain adaptability.
- We propose the co-adaptive prototypical networks to perform co-adaptive feature transformation through the transformer mechanism, which helps to obtain distinguishable prototype and query representations.
- Experimental results on the benchmark datasets show that our approach significantly outperforms the baseline models and achieves state-of-the-art DAFSRE performance.

2. Task definition

DAFSRE aims to predict novel classes using only a few labeled instances, where the domain of the test set differs from the training set. The N-way-K-shot setting is widely used to simulate low-resource RE scenarios. This setting includes a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. The support set $\mathcal{S} = \{s_k^i; i = 1, ..., N, k = 1, ..., K\}$ consists of N novel classes, each with a small number of labeled instances K. The instances in $\mathcal{Q} = \{q_j; j = 1, ..., R\}$ need to be classified using the given $N \times K$ support instances. An auxiliary dataset $\mathcal{D}_{train}$ is provided to train the encoder, which contains abundant general classes, each with a large number of labeled instances. Each instance in $\mathcal{S}$, $\mathcal{Q}$ and $\mathcal{D}_{train}$ contains a sentence with pre-annotated entities. It is important to note that source and target domain relation classes are disjoint to ensure the few-shot learning scenarios. Furthermore, the unlabeled data from the target domain $\mathcal{T}$ is provided, which may enable us to explore some pertinent information specific to the target domain. The goal of DAFSRE is to optimize the following objective function:

$$\mathcal{L} = -\frac{1}{R}\sum_{q\in\mathcal{Q}} P(y_q \mid \mathcal{S}, q). \tag{1}$$

3. Methodology

The overall framework is illustrated in Figure 2. DCFT mainly consists of three components: Encoder, Domain-aware Transformation Module (DTM) and Co-adaptive Prototypical Networks (CPN). In this section, we will present the details of our proposed DCFT approach.

3.1. Encoder

A good feature representation is crucial for solving tasks (Yao et al., 2023b; Chen et al., 2023a; Yao et al., 2023a). The encoder is used to convert each word in a sentence into a contextualized embedding. Consistent with prior research (Han et al., 2021a,b), we adopt BERT (Devlin et al., 2019) as the sentence encoder. Considering the vital role of entity words in DAFSRE tasks, we enclose entities within the sentence with special tokens [E] and [\E] to mark the entity boundary and emphasize the entity for the encoder. After providing the sentence's tokens to BERT, we can obtain the embedding of each token.

$$\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_m = Encoder(x), \mathbf{h}_i \in \mathbb{R}^{d_e}, \tag{2}$$

where d_e is the dimension of the entity embedding.

Then, we concatenate the head and tail entity embeddings $\mathbf{h}_{head}, \mathbf{h}_{tail}$, as well as the difference between them $\mathbf{h}_{head} - \mathbf{h}_{tail}$ to obtain the representation of each support or query instance $\mathbf{h}^{'}$.

$$\mathbf{h}^{'} = [\mathbf{h}_{head}; \mathbf{h}_{tail}; \mathbf{h}_{head} - \mathbf{h}_{tail}], \mathbf{h}^{'} \in \mathbb{R}^{1\times 3de}. \tag{3}$$

3.2. Domain-aware transformation module

Our focus in this module is to enhance the adaptability of the model in a specific target domain. As illustrated in Figure 1(a), the semantic gap between the source and target domain can sometimes be considerable. This may result in the model becoming overfitted to the source domain and faltering in generalizing to the target domain. To tackle this challenge, we aim to simulate the feature distributions of the target domain during the training phase through explicit target-domain-guided feature transformation. By providing the model with more information about the feature distribution of the target domain, we can explicitly guide the feature transformation that makes source domain data more

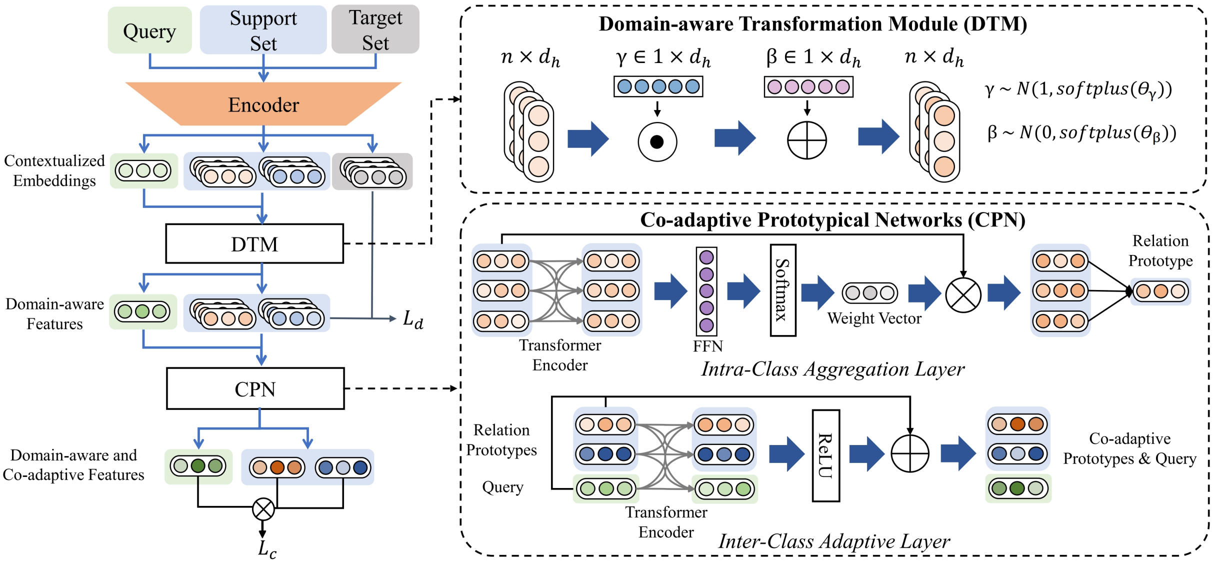

Query
Support Set
Target Set
Encoder
Contextualized Embeddings
DTM
Domain-aware Features
L_d
CPN
Domain-aware and Co-adaptive Features
L_c
Domain-aware Transformation Module (DTM)
$n \times d_h$
$\gamma \in 1 \times d_h$
$\beta \in 1 \times d_h$
$n \times d_h$
$\gamma \sim N(1, softplus(\theta_\gamma))$
$\beta \sim N(0, softplus(\theta_\beta))$
Co-adaptive Prototypical Networks (CPN)
Transformer Encoder
FFN
Softmax
Weight Vector
Relation Prototype
Intra-Class Aggregation Layer
Relation Prototypes
Query
Transformer Encoder
ReLU
Co-adaptive Prototypes & Query
Inter-Class Adaptive Layer

Figure 2: The framework of the proposed model DCFT. $\mathcal{L}_d$ is the domain discrepancy loss and $\mathcal{L}_c$ is the classification loss.

closely aligned with the target domain as shown in Figure 1(b).

We achieve this goal by a learnable domain-aware feature transformation layer with domain discrepancy loss. Our approach is inspired by recent research indicating that the mean and standard deviation of features encode essential "domain style" information, making them suitable for domain-aware transformation and manipulation in a reliable and effective manner (Li et al., 2021; Tang et al., 2021). Specifically, as shown in Figure 2, We introduce a domain-aware feature transformation layer that follows the encoder. This layer is governed by hyper-parameters $\theta_\gamma \in \mathbb{R}^{1\times 3de}$ and $\theta_\beta \in \mathbb{R}^{1\times 3de}$, which represent the standard deviations of the Gaussian distributions used to sample the scaling and bias terms for the affine transformation, respectively. To transform an instance's representation $\mathbf{h}^{'}$, which has a dimension of $1 \times 3d_e$, we first randomly sample the scaling term γ and bias term β from Gaussian distributions $N(\cdot)$,

$$\gamma \sim N\left(\mathbf{1}, \mathrm{softplus}\left(\theta_\gamma\right)\right), \beta \sim N\left(\mathbf{0}, \mathrm{softplus}\left(\theta_\beta\right)\right). \tag{4}$$

Then we compute the domain-aware feature as:

$$\hat{\mathbf{h}} = \gamma \times \mathbf{h}^{'} + \beta. \tag{5}$$

Although it is possible to empirically determine hyper-parameters θ_γ and θ_β of the domain-aware feature transformation layer, manually adjusting a universal set of parameters that perform well across various scenarios remains difficult. So we design a domain discrepancy loss that exploits the "domain style" information we mentioned before. This loss provides supervised information for the transformation process and dynamically adjusts the θ_γ and θ_β. It is computed as the difference in mean and standard deviation between the domain-aware features and the contextualized embeddings of some target domain instances. By utilizing this loss, the domain-aware feature transformation layer is encouraged to enhance features that are more aligned with the target domain. As a result, the model can effectively train on the data related to target domain features and improve its performance on the target domain. The domain discrepancy loss is calculated by the following formula:

$$\begin{aligned}\mathcal{L}_d = &(\mathrm{mean}(\mathrm{DTM}(\mathcal{S})) - \mathrm{mean}(\mathrm{DTM}(\mathcal{T}_{rs})))^2 \\ &+\eta(\mathrm{std}(\mathrm{DTM}(\mathcal{S})) - \mathrm{std}(\mathrm{DTM}(\mathcal{T}_{rs})))^2,\end{aligned} \tag{6}$$

where $\mathrm{DTM}(\cdot)$ indicates the domain-aware feature transformation operation as mentioned previously, $\mathrm{mean}(\cdot)$ and $\mathrm{std}(\cdot)$ are functions used to calculate the mean and standard deviation of a set of feature vectors, respectively. $\mathcal{T}_{rs}$ is a subset of $\mathcal{T}$ obtained by random sampling. η is used for balancing the weight of mean and standard deviation discrepancy.

3.3. Co-adaptive prototypical networks

To facilitate the co-adaptation of domain-aware support and query features, we propose the co-adaptive prototypical networks (CPN), which employ a self-attention mechanism to further transform and improve the features. The CPN is composed of two transformation layers: the intra-class aggregation layer and the inter-class adaptive layer.

Intra-class aggregation layer: In the original prototypical networks, the relation prototype is obtained by taking the average of the features of all instances within a relation. However, in the case of few-shot scenarios, the absence of sufficient supporting data can result in a significant deviation of the prototype if there is an instance that the representation differs greatly from that of the others. This is especially prevalent in situations where the data is noisy or where the relations exhibit a wide range of semantics. Such phenomena lead to unsuitable prototypes that can adversely impact the accuracy of classification.

To improve the effectiveness of prototypical networks, we propose the intra-class aggregation layer. We argue that each instance within a relation should have a distinct weight, and that this weight should be determined based on all instances within that relation. The proposed layer would give priority to the more informative and central instances that are better representative of the relation prototype. Specifically, given the features of all instances of the same relation after the DMT module, we first employ the Transformer Encoder (TE) layer (Vaswani et al., 2017) to these features to capture their intra-class interactions. It is based on the self-attention function $\mathrm{Self\text{-}Att}(\cdot)$:

$$\mathrm{Self\text{-}Att}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{3de}}\right)\mathbf{V}, \quad (7)$$

where $\mathbf{Q}=\mathbf{K}=\mathbf{V}\in\mathbb{R}^{k\times 3de}$ is the domain-aware features of all instances within a relation. $\sqrt{3de}$ is used to rescale the inner product of two vectors. This formulation was also extended to a multi-head self-attention (MHSA). The TE is combined with the MHSA, feed-forward network (FFN), residual connection, and layer normalization:

$$\mathrm{head_i} = \mathrm{Self\text{-}Att}\left(\mathbf{Q}W_i^{\mathbf{Q}}, \mathbf{K}W_i^{\mathbf{K}}, \mathbf{V}W_i^{\mathbf{V}}\right), \quad (8)$$

$$\mathrm{MHSA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Mean}\left(\mathrm{head_1},\ldots,\mathrm{head_h}\right), \quad (9)$$

$$\mathbf{V}^0 = \mathrm{MHSA}(\mathbf{Q},\mathbf{K},\mathbf{V}) + \mathbf{V}, \quad (10)$$

$$\mathrm{TE}(\mathbf{V}) = \Phi(\mathbf{V}^0 + \mathrm{FFN}(\Phi(\mathbf{V}^0))), \quad (11)$$

where $W_i^{\mathbf{Q}}, W_i^{\mathbf{K}}, W_i^{\mathbf{V}} \in \mathbb{R}^{3de\times 3de}$ is the learnable feature projection matrix of i-th head for $\mathbf{Q},\mathbf{K}$ and $\mathbf{V}$. $\mathbf{V}^0$ is the output of multi-head self-attention with residual connection. Φ indicates the normalization layer.

After we get the features processed by $\mathrm{TE}(\cdot)$, we put them into an FFN and a softmax layer to map the features of each instance to a one-dimensional scalar, then we compute the weight of each instance:

$$\alpha^i = \mathrm{Softmax}(\mathrm{FFN}(\hat{\mathbf{S}}^{i\prime})), \quad (12)$$

where $\alpha^i = \{\alpha_k^i; k = 1,2,\ldots,K\}$ is the weight vector for the instances in relation i. $\hat{\mathbf{S}}^{i\prime}$ indicates the all instances features in relation i processed by $\mathrm{DTM}(\cdot)$ and $\mathrm{TE}(\cdot)$.

Then, we compute the weighted sum of the domain-aware features within a relation i with α^i to obtain the corresponding intra-class adaptive relation prototype $\mathbf{p}^i$.

$$\mathbf{p}^i = \sum_k^K \alpha_k^i \hat{\mathbf{s}}_k^i, \quad (13)$$

where $\hat{\mathbf{s}}_k^i$ indicates the domain-aware feature of k-th instance in relation i.

Inter-class adaptive layer: Once we have obtained a prototype for each relation, we can simply calculate the distances between the query and each prototype to perform classification based on the similarity score. However, this approach fails to account for discriminative information among different relation prototypes, which may significantly impair performance when dealing with confusing relations.

To capture the complex interactions among all relation prototypes and query instances, we propose the inter-class adaptive layer. By employing a TE layer in the relation prototypes and query features, we can focus more on the discriminative feature dimensions that aid in differentiating between various relation prototypes. The query features can also be improved by attending to relevant features for each relation prototype, thereby enhancing the classification performance. Concretely, given each intra-class adaptive relation prototype $\mathbf{p}^i$ and the domain-aware query $\hat{\mathbf{q}}$, we first put them into the TE and an activation function to obtain the discriminative features:

$$[\dot{\mathbf{p}}^1,\ldots,\dot{\mathbf{p}}^N,\dot{\mathbf{q}}] = \sigma(\mathrm{TE}([\mathbf{p}^1,\ldots,\mathbf{p}^N,\hat{\mathbf{q}}])), \quad (14)$$

where σ represents the ReLU function, $\dot{\mathbf{p}}^i$ stands for the discriminative feature of relation i, $\dot{\mathbf{q}}^i$ stands for the discriminative feature of query instance.

Then, we perform a residual connection to make the relation prototypes and query pay more attention to these discriminative features:

$$[\tilde{\mathbf{p}}^1,\ldots,\tilde{\mathbf{p}}^N,\tilde{\mathbf{q}}] = [\dot{\mathbf{p}}^1,\ldots,\dot{\mathbf{p}}^N,\dot{\mathbf{q}}] \oplus [\mathbf{p}^1,\ldots,\mathbf{p}^N,\hat{\mathbf{q}}], \quad (15)$$

where $\oplus$ is the element-wise plus function, $\tilde{\mathbf{p}}^i$ and $\tilde{\mathbf{q}}$ indicates the final domain-aware and co-adaptive features of relation prototypes and the query.

Finally, the label probability of query for classification is:

$$p(y=r \mid \mathcal{S}, q) = \frac{\exp\left(d\left(\tilde{\mathbf{p}}^r, \tilde{\mathbf{q}}\right)\right)}{\sum_{i=1}^N \exp\left(d\left(\tilde{\mathbf{p}}^i, \tilde{\mathbf{q}}\right)\right)}, \quad (16)$$

where $d(\cdot)$ indicates the squared Euclidean distance between two vectors.

Table 2: Accuracy (%) of models on FewRel 2.0 Pubmed test set under N-way-K-shot (NW-KS) settings. **Best** (bold) and second best numbers are highlighted in each column. We run all the algorithms on the same conditions.

Model	5W-1S	5W-5S	10W-1S	10W-5S	AVG
Proto (Snell et al., 2017)	40.12	51.50	26.45	36.93	38.75
Proto-ADV (Snell et al., 2017)	41.90	54.74	27.36	37.40	40.35
MAML (Finn et al., 2017)	66.62	78.53	51.90	65.57	65.66
BERT-PAIR (Gao et al., 2019)	67.41	78.57	54.89	66.85	66.93
DaFeC (Cong et al., 2021)	61.20	76.99	47.63	64.79	62.65
REGRAB (Qu et al., 2020)	71.70	80.74	61.66	74.06	72.04
REGRAB-ADV (Qu et al., 2020)	65.10	71.61	56.44	56.71	62.47
MTB (Soares et al., 2019)	74.7	87.9	62.5	81.1	76.6
CP (Peng et al., 2020)	79.7	84.9	68.1	79.8	78.1
HCPR (Han et al., 2021a)	76.34	83.03	63.77	72.94	74.02
FAFE (Dou et al., 2022)	73.58	90.10	62.98	80.51	76.79
GM_GEN (Li and Qian, 2022)	76.67	91.28	64.19	84.8	79.24
ChatGPT (OpenAI, 2023)	66.34	78.06	55.19	65.81	66.35
DCFT	**82.54**	**93.27**	**71.44**	**87.98**	**83.81**

In our approach, we have adopted the cross-entropy loss as the classification loss function. It can be defined as follows:

$$\mathcal{L}_c(\mathcal{S}, q) = -\log p(y = t \mid \mathcal{S}, q), \tag{17}$$

where t stands for ground truth label.

The final objective function of our model is defined as $\mathcal{L} = \mathcal{L}_c + \lambda\mathcal{L}_d$, where λ is a hyperparameter used to balance the two terms.

4. Experiments

4.1. Dataset and evaluation

Dataset: We conduct experiments on the FewRel 2.0 dataset (Gao et al., 2019) since it is the **only qualified large-scale** dataset suitable for evaluating the DAFSRE model. This dataset comprises 4 sub-datasets from four different domains, including Wikipedia, SemEval-2010 task8, NYT, and Pubmed. In our experimental setup, we take 64/10/15 relations, with a total of 44800/1000/1500 labeled instances for training/validation/testing. The training set is drawn from the humanities domain, and constructed using articles from Wikipedia. In contrast, the standard validation and test sets derived from the medical domain, aligned through the matching of PubMed[1] with UMLS[2]. Additionally, we used the simulated-Semeval dataset to conduct another experiment. In our experiments, the Wikipedia data served as the source domain, while the Pubmed and simulated-Semeval data were the target domains.

Evaluation: The standard metric used to evaluate the performance of the DAFSRE model is the average accuracy of the N-way-K-shot task. According to the previous works (Gao et al., 2019; Han et al., 2018) , we have set N to 5 and 10 and K to 1 and 5, thus forming four few-shot learning scenarios. As the test set labels in FewRel 2.0 are not publicly available, we report the final test accuracy by submitting the model's predictions to the FewRel Leaderboard[3].

4.2. Implementation details

Our approach is implemented with PyTorch (Paszke et al., 2019). We take the uncased model of BERT$_{base}$ as the encoder in our approach. In our training phase, we set the number of iterations to 5000, and perform validation every 200 iterations. The learning rate is set to 2e-5, while λ is set to 0.5, and η is set to e-3. We train and evaluate our model using one GeForce RTX 3090 GPU with about 24 GB of memory.

4.3. Comparison with previous works

Tables 2 and Tables 3 compare the proposed approach with current state-of-the-art methods. In addition to **ChatGPT**[4], we divide other models into two categories based on whether they introduce additional information (e.g. relation description information, additional large-scale pre-training data), including: **Proto** (Snell et al., 2017), **Proto-ADV** (Snell et al., 2017), **MAML** (Finn et al., 2017),

[1] https://www.ncbi.nlm.nih.gov/pubmed/

[2] https://www.nlm.nih.gov/research/umls/

[3] https://thunlp.github.io/fewrel.html

[4] The version of ChatGPT we used is gpt-3.5-turbo-0613

Table 3: Accuracy (%) of models on simulated-Semeval dataset under N-way-K-shot (NW-KS) settings. * is our re-implementation based on the corresponding open-source codes. $\star$ is reported by Yuan et al. (Yuan et al., 2023)

Model	5W-1S	5W-5S	10W-1S	10W-5S	AVG
Proto* (Snell et al., 2017)	48.54	78.48	36.43	68.15	57.90
Proto-ADV* (Snell et al., 2017)	39.79	58.76	27.23	45.46	42.81
MAML$\star$ (Finn et al., 2017)	42.75	52.87	27.89	43.06	41.64
BERT-PAIR* (Gao et al., 2019)	49.70	67.64	37.71	55.14	52.55
DaFeC* (Cong et al., 2021)	48.04	57.07	34.36	42.66	45.53
REGRAB$\star$ (Qu et al., 2020)	49.56	64.57	36.17	54.10	51.10
REGRAB-ADV$\star$ (Qu et al., 2020)	50.71	65.46	38.61	54.56	52.34
HCPR* (Han et al., 2021a)	56.98	73.65	43.75	62.64	59.26
FAFE* (Dou et al., 2022)	59.03	76.99	46.27	66.48	62.19
GM_GEN* (Li and Qian, 2022)	51.48	79.02	44.96	69.86	61.33
ChatGPT (OpenAI, 2023)	46.75	67.17	38.85	52.43	51.30
DCFT	**59.46**	**80.31**	**47.21**	**72.13**	**64.78**

BERT-PAIR (Gao et al., 2019), **DaFeC** (Cong et al., 2021), **REGRAB** (Qu et al., 2020), **REGRAB-ADV** (Qu et al., 2020). And external information enhanced models: **MTB** (Soares et al., 2019), **CP** (Peng et al., 2020), **HCPR** (Han et al., 2021a), **FAEA** (Dou et al., 2022), **GM_GEN** (Li and Qian, 2022). The details of these models are presented in the Related work section. As for ChatGPT, during inference phase, we follow the (Brown et al., 2020) to construct prompts, which concatenates a few demonstrations and the test sample to predict the label.

From the tables, we can observe that: 1) Our approach consistently outperforms most methods across all settings. Notably, our approach significantly improves the accuracy by an average of 4.57% compared to the second-best model GM_GEN on FewRel 2.0 Pubmed test set. On another dataset, the simulated-Semeval dataset, DCFT also yields gains of 2.59% compared to the second-best model FAFE. It demonstrates the effectiveness of DCFT in addressing the DAFSRE task and achieving state-of-the-art performance. 2) Compared to methods like MTB and CP, **which rely on extensive pre-training with external data**, (i.e., about 600 million relation statement sentences, much higher than the 44,800 sentences of the FewRel 2.0 dataset), our approach still improved the average accuracy by over 5%. This indicates that our approach has strong domain adaptability and few-shot learning ability, even without pre-training on large amounts of relation extraction data. 3) As for the LLMs, their performance in this task is unexpectedly poor. This may be due to their limited understanding in medical domain, which can lead to model hallucinations (Zhang et al., 2023b) and they confidently output incorrect answers. Moreover, the performance of the LLMs can also be compromised by excessively long text inputs (e.g. 10-way-5-shot scenario) due to catastrophic forgetting.

Table 4: Experimental results of the ablation studies on the FewRel 2.0 PubMed test set.

Model	5W-1S	5W-5S	AVG
DCFT	**82.54**	**93.27**	**87.91**
w/o DTM	79.36	90.71	85.04
w/o CPN	80.63	90.15	85.39
w/o CPN-Intra	82.54	91.32	86.93
w/o CPN-Inter	80.63	91.69	86.16

4.4. Ablation study

In this section, we remove specific components or modules on the model to perform ablation study. In Table 4, the w/o DTM means the model without DTM, i.e., the instance feature obtained by the encoder was directly fed into CPN for classification. The w/o CPN indicates the model without CPN, the instances within a relation are directly averaged to calculate the prototype for classification. We also disassemble the CPN module to investigate the contributions of different components.

Effect of DTM: The results of w/o DMT are presented in Table 4, where each experimental configuration exhibits a decline to a certain extent, with an average reduction of 2.87%. This table reveals that the DTM holds more significance in the 1-shot scenario as the accuracy dropped by 3.18%. This finding highlights the crucial role of the target domain-guided transformation in DAFSRE, particularly when the available data is scarce.

Effect of CPN: According to Table 4, the absence of the CPN results in a decline in accuracy by 1.91%

Table 5: Experimental results of the variants of our model on the FewRel 2.0 PubMed test set.

Model	5W-1S	5W-5S	AVG
DCFT	**82.54**	**93.27**	**87.91**
DCFT-fixed	80.84	92.78	86.81
DCFT-random	81.09	90.46	85.78
DCFT-ADV	79.82	92.15	85.99
-BERT-large	81.62	91.88	86.75
-RoBERTa-base	81.11	93.22	87.17
-RoBERTa-large	80.22	92.54	86.38

and 3.12% in 1-shot and 5-shot scenarios, respectively. This observation implies that the inter-class aggregation layer and inter-class adaptive layer are able to capture interactive information and improve the query and prototype representations, leading to better overall model performance.

4.5. Variations experiments

Additionally, we intend to explore some other variants of our model to gain a more comprehensive understanding of its capabilities. As shown in Table 4, DCFT-fixed fixes the hyper-parameters θ_γ and θ_β in DTM at each step and eliminate the domain discrepancy loss, while DCFT-random randomizes these hyper-parameters. Another variation is DCFT-ADV, which replaces the DTM with a domain-adversarial part following the approach of DANN (Ganin et al., 2016). -BERT-large, -RoBERTa-base and -RoBERTa-large indicate using BERT and RoBERTa(Liu et al., 2019) models of different scales as encoders.

Module variations: Table 5 shows that all variations in module show a significant decline in accuracy, especially the domain-adversarial training version DCFT-ADV. Notably, DCFT-random yields an improvement (+1.73%) in the 1-shot scenario compared to the w/o DMT ablation. However, it undermines the model's performance (-0.25%) in the 5-shot scenario, indicating the instability of this variant. Furthermore, DCFT-fixed shows unstable improvement compared to the w/o DMT ablation as well. These findings underscore the crucial role of unsupervised target domain data in guiding feature transformation, which can enhance the model's cross-domain adaptability in a stable manner.

Backbone variations: To verify the effectiveness of our method on different backbones, we replaced the BERT-base encoder with BERT-large, RoBERTa-base, and RoBERTa-large, respectively. The results show that our method performs well on these variations. However, from Table 5, we can find that replacing the larger-scale encoders does not significantly improve model performance. This is mainly due to the limited training data, which prevents sufficient training of larger models.

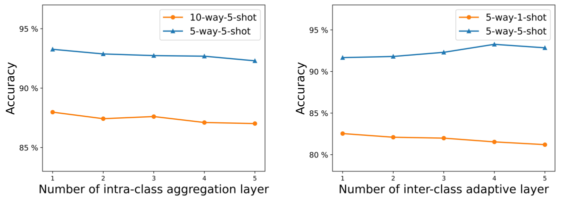


Figure 3: Impacts of the number of intra-class aggregation layer (left) and inter-class adaptive layer (right) in different few-shot scenarios on FewRel 2.0 Pubmed test set.

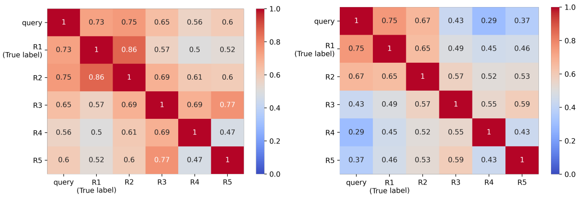


Figure 4: A example of correlation scores between query and relation prototypes without (left) and with (right) CPN. A higher score indicates a higher similarity of prototypes, making it hard to distinguish.

4.6. Parameter Analysis

To explore the impact of number of intra-class aggregation layer and inter-class adaptive layer in different few-shot scenarios, we conduct experiments with different layer number as reported in Figure 3. Specifically, one intra-class aggregation layer and four inter-class adaptive layer can achieve optimal performance in 5-shot scenarios. While one inter-class adaptive layer can achieve best performance in 1-shot scenario. This indicates that a single intra-class aggregation layer is sufficient to model the interaction between instances well. When there are fewer instances, excessive inter-class adaptive layers may lead to overfitting of the model and damage its performance.

4.7. Visualization

In Fig 4, we present the correlation score between query and relation prototypes with and without CPN. Without CPN, the correlation score between the query and prototypes was found to be high, indicating that the prototypes were more similar to each other, making it difficult to classify the query. However, with the help of CPN, the model is able to generate more distinguishable features for the relation prototypes, resulting in a lower similarity score among them and ultimately enhancing the model's classification performance.

5. Related work

Early works usually apply metric learning-based methods (Koch et al., 2015; Qu et al., 2020; Han et al., 2021a) to resolve DAFSRE, which leverage the distance distribution of each relation to perform classification. For example, DaFeC (Cong et al., 2021) applies a clustering promotion mechanism to learn better features for the target domain. BERR-PAIR (Gao et al., 2019) measures the similarity of sentence pairs to classify instances and achieve impressive performance. More recently, meta-learning-based approaches, such as MAML (Finn et al., 2017) and Proto (Snell et al., 2017), have become popular in DAFSRE tasks. REGRAB (Qu et al., 2020) completes the FSRE task via Bayesian meta-learning on the relation graph. REGRAB-ADV (Qu et al., 2020) adds an adversarial part to the REGRAB model. Since the information available in a single sentence is limited in few-shot learning scenarios, recent works such as MTB (Soares et al., 2019) and CP (Peng et al., 2020) have taken an extensive pre-training approach with external data. Furthermore, HCPR (Han et al., 2021a) proposes a hybrid contrastive relation-prototype approach that focuses on hard few-shot relation extraction tasks. FEFA (Dou et al., 2022) utilizes a function words adaptively enhanced attention framework to attend to class-related function words with relation descriptions, achieving impressive results. GM_GEN (Li and Qian, 2022) proposes to generate a general model for all tasks and finetune to get tiny task-specific models. Recently, LLMs like ChatGPT (OpenAI, 2023) have been pre-trained on extensive datas, which exhibit remarkable reasoning capabilities across many NLP tasks. However, despite the significant progress in DAFSRE, the current methods exhibit limitations in effectively leveraging the feature distribution of the target domain and considering the interaction between instances.

6. Conclusion

In this paper, we propose the domain-aware and co-adaptive feature transformation approach, which aims to mitigate the domain gap and facilitate learning to distinguish confusing relations for DAFSRE. We introduce a domain-aware transformation module that conducts target-domain-guided feature transformation, which enhances the model's adaptability to the target domain. Additionally, the co-adaptive prototypical networks are utilized to model both intra- and inter-class interactions among all instances, leading to improved classification performance. Experimental results demonstrate the effectiveness of our approach and achieve state-of-the-art performance on the DAFSRE benchmark.

7. Acknowledgement

This work was supported by No.YJS003HGC1.

8. Bibliographical References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jushuo Chen, Feifei Dai, Xiaoyan Gu, Haihui Fan, Jiang Zhou, Bo Li, and Weiping Wang. 2023a. Learning pair-centric representation for link sign prediction with subgraph. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 256–265.

Jushuo Chen, Feifei Dai, Xiaoyan Gu, Jiang Zhou, Bo Li, and Weipinng Wang. 2023b. Universal domain adaptive network embedding for node classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4022–4030.

Xin Cong, Bowen Yu, Tingwen Liu, Shiyao Cui, Hengzhu Tang, and Bin Wang. 2021. Inductive unsupervised domain adaptation for few-shot classification via clustering. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part II*, pages 624–639. Springer.

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

Chunliu Dou, Shaojuan Wu, Xiaowang Zhang, Zhiyong Feng, and Kewen Wang. 2022. Function-words adaptively enhanced attention networks for few-shot inverse relation classification. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 2937–2943.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Jiale Han, Bo Cheng, and Wei Lu. 2021a. Exploring task difficulty for few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2616.

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 *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 649–658.

Yanxu Hu and Andy J Ma. 2022. Adversarial feature augmentation for cross-domain few-shot classification. In *European Conference on Computer Vision*, pages 20–37. Springer.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *Proceedings of ICML workshop on deep learning*, volume 2, page 0. Lille.

Boyi Li, Felix Wu, Ser-Nam Lim, Serge Belongie, and Kilian Q Weinberger. 2021. On feature normalization and data augmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12378–12387. IEEE.

Qingyao Li, Hui Xu, Hui Wang, and Buzhou Tang. 2022a. S3 aal: Support set selection based on adversarial active learning for medical few-shot relation extraction. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 777–780. IEEE.

Wanli Li and Tieyun Qian. 2022. Graph-based model generation for few-shot relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 62–71.

Yile Li, Yijun Liu, GU Xiaoyan, Yinliang Yue, Haihui Fan, and Bo Li. 2022b. Dual reasoning based pairwise representation network for document level relation extraction. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Yijun Liu, Feifei Dai, Xiaoyan Gu, Haihui Fan, Dong Liu, Bo Li, and Weiping Wang. 2023. Powering fine-tuning: Learning compatible and class-sensitive representations for domain adaption few-shot relation extraction. In *International Conference on Database Systems for Advanced Applications*, pages 121–131. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

OpenAI. 2023. Openai official website. introducing chatgpt: https://openai.com/blog/chatgpt.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3661–3672.

Meng Qu, Tianyu Gao, Louis-Pascal Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *International conference on machine learning*, pages 7867–7876. PMLR.

Afia Sajeeda and BM Mainul Hossain. 2022. Exploring generative adversarial networks and adversarial training. *International Journal of Cognitive Computing in Engineering*, 3:78–89.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Livio Baldini Soares, Nicholas Fitzgerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.

Zhiqiang Tang, Yunhe Gao, Yi Zhu, Zhi Zhang, Mu Li, and Dimitris N Metaxas. 2021. Crossnorm and selfnorm for generalization under distribution shifts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 52–61.

Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. 2019. Cross-domain few-shot classification via learned feature-wise transformation. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Suhe Wang, Xiaoyuan Liu, Bo Liu, and Diwen Dong. 2022. Sentence-aware adversarial meta-learning for few-shot text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4844–4852.

Jiawei Yao and Yingxin Lai. 2023. Dynamicbev: Leveraging dynamic queries and temporal context for 3d object detection. *arXiv preprint arXiv:2310.05989*.

Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. 2023a. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9455–9465.

Jiawei Yao, Tong Wu, and Xiaofeng Zhang. 2023b. Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with cnn. *arXiv preprint arXiv:2308.08333*.

Zhongju Yuan, Zhenkun Wang, and Genghui Li. 2023. Cross-domain few-shot relation extraction via representation learning and domain adaptation. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Minghui Zhai, Feifei Dai, Xiaoyan Gu, Haihui Fan, Dong Liu, and Bo Li. 2023. Learning discriminative semantic and multi-view context for domain adaptive few-shot relation extraction. In *International Conference on Neural Information Processing*, pages 283–296. Springer.

Ji Zhang, Jingkuan Song, Lianli Gao, and Hengtao Shen. 2022. Free-lunch for cross-domain few-shot learning: Style-aware episodic training with robust contrastive learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2586–2594.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023a. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Xingyu Zhu, Feifei Dai, Xiaoyan Gu, Haihui Fan, Bo Li, and Weiping Wang. 2023. Erpg: Enhancing entity representations with prompt guidance for complex named entity recognition. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2813–2818. IEEE.

9. Language Resource References

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. 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 (EMNLP-IJCNLP)*, pages 6250–6255.

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*, pages 4803–4809.