Knowledge-Enhanced Domain Adaptation in Few-Shot Relation Classification

Jiawen Zhang
Institute of Software, Chinese Academy of Sciences, University of CAS, Beijing, China
Zhejiang Lab, Hangzhou, China
jiawen2019@iscas.ac.cn

Jiaqi Zhu*
Institute of Software, Chinese Academy of Sciences, University of CAS, Beijing, China
Zhejiang Lab, Hangzhou, China
zhujq@ios.ac.cn

Yi Yang
Institute of Software, Chinese Academy of Sciences, University of CAS, Beijing, China
yangyi2012@iscas.ac.cn

Wandong Shi
Institute of Software, Chinese Academy of Sciences, University of CAS, Beijing, China

Congcong Zhang
Institute of Software, Chinese Academy of Sciences, University of CAS, Beijing, China

Hongan Wang
Institute of Software, Chinese Academy of Sciences, University of CAS, Beijing, China

ABSTRACT

Relation classification (RC) is an important task in knowledge extraction from texts, while data-driven approaches, although achieving high performance, heavily rely on a large amount of annotated training data. Recently, many few-shot RC models have been proposed and yielded promising results in general domain datasets, but when adapting to a specific domain, such as medicine, the performance drops dramatically. In this paper, we propose a **K**nowledge-**E**nhanced **F**ew-shot RC model for the **D**omain **A**daptation task (KEFDA), which incorporates general and domain-specific knowledge graphs (KGs) to the RC model to improve its domain adaptability. With the help of concept-level KGs, the model can better understand the semantics of texts and easily summarize the global semantics of relation types from only a few instances. To be more important, as a kind of meta-information, the manner of utilizing KGs can be transferred from existing tasks to new tasks, even across domains. Specifically, we design a knowledge-enhanced prototypical network to conduct instance matching, and a relation-meta learning network for implicit relation matching. The two scoring functions are combined to infer the relation type of a new instance. Experimental results on the Domain Adaptation Challenge in the *FewRel 2.0* benchmark demonstrate that our approach significantly outperforms the state-of-the-art models (by 6.63% on average).[1]

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Transfer learning**; *Semantic networks.*

KEYWORDS

relation classification; few-shot learning; knowledge graph; domain adaptation; relation meta

ACM Reference Format:
Jiawen Zhang, Jiaqi Zhu, Yi Yang, Wandong Shi, Congcong Zhang, and Hongan Wang. 2021. Knowledge-Enhanced Domain Adaptation in Few-Shot Relation Classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548.3467438

1 INTRODUCTION

Discovering knowledge from unstructured data is always a significant but difficult topic, especially for professional areas where knowledge is updated and evolving over time. For example, to keep up with the latest advances in the medical field, doctors need to read up-to-date medical literature to find useful information, which is time-consuming. For this reason, increasing works are trying to obtain specialized knowledge automatically from plain texts.

Relation classification (RC) [3] is an important task in knowledge extraction from texts. Given a sentence, RC aims at determining the relation type between two entities based on the contextual semantic information. Currently, the main successful research for RC includes rule-based approaches [12, 14] and data-driven approaches [4, 16, 39, 41]. However, the former are generally ad-hoc and vulnerable to the error accumulation issue, and the latter heavily depend on annotated training data. For professional areas, such as medicine and finance, sufficient domain-specific extraction rules and labeled data are harder to access due to the required expertise.

More recently, many researchers began to regard the RC task as a few-shot learning (FSL) problem [33], and solve it through rapidly transferring some meta information from existing tasks to new tasks containing only a few samples with supervised information [5, 29, 34]. Although these models have yielded promising results in the general domain, their performance drops dramatically when adapting to a new specific domain due to the discrepancy in morphology, syntax, and semantics across domains.

Intuitively, in addition to the ability of natural language processing (NLP), the background knowledge of this field is also crucial for the domain-specific text understanding. Fortunately, a lot of

*Corresponding Author
[1]The source code is publicly available at https://github.com/imJiawen/KEFDA.

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467438

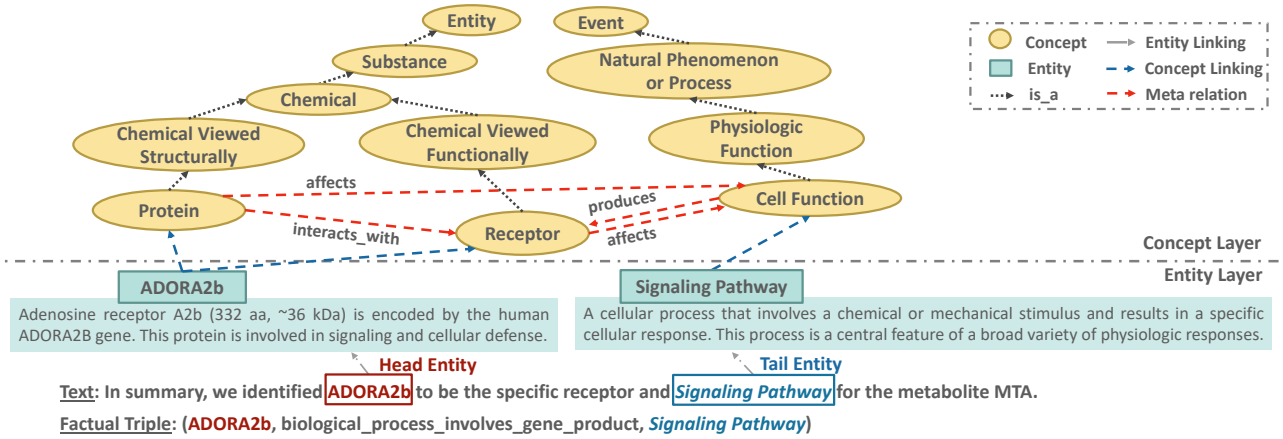

Figure 1: Example of a medical text, the factual triple it contains, and the relevant part of a medical KG, *UMLS*. The KG provides rich semantic and descriptive information of the head and tail entities, which is beneficial for understanding the text.

knowledge graphs (KGs) on specific domains have been constructed, published, and employed in various scenarios, such as medical KG[2], cultural heritage KG[3], academic KG[4] and e-commerce KG[5]. Inspired by this, we develop the idea of utilizing the domain-specific KG in the RC model to improve its adaptability.

It is worth noting that many efforts have been devoted to fusing existing KGs into NLP models in the pre-training phase [18, 27, 32, 40], but these works have several common limitations and are not suitable for our task. To begin with, since only the general domain KG is used, when adapting to a specific domain, such prior knowledge no longer provides significant and targeted background information. If we pre-train the model again with the domain-specific KG and corpus for each domain, it would be very time-consuming and laborious, let alone such resources are not always available. Ideally, the domain-specific KG should be applied directly to downstream tasks.

Moreover, all of existing approaches utilize entity-level KGs, which suffer from the problem of quick knowledge update, large storage space, and heavy computational efforts. Comparatively, we believe that the lightweight concept-level KG (i.e. ontology) is a better choice. From the perspective of text understanding, implicit relations in the concept layer are more stable and can provide richer semantic information to the texts. For example, as shown in Figure 1, with the help of domain-specific KG, we can know that the head entity (`ADORA2b`) of this sentence belongs to a `receptor` and the tail entity (`signaling pathway`) is of type `cell function`, and the factual triple (`receptor, affect, cell function`) in the concept-level KG makes us understand the text easier. In addition to the semantic enhancement on individual instances, the concept-level KG can even help us summarize the semantics of relation types with a small set of instances in the global view, and better infer the relation type of a new instance. Furthermore, from the perspective of domain adaptation (DA), the manner of utilizing KGs above can be treated as a kind of meta-information, which is able to be transferred from existing tasks in one domain to new tasks in another domain, since the concept-level KGs in different domains have more structural commonalities than entity-level KGs.

In this paper, we propose KEFDA, a novel model for knowledge-enhanced domain adaptation in the few-shot RC task, which directly employs the domain-specific KG, especially its concept-level part, in the downstream task. To achieve this, a series of technical challenges are raised and will be tackled in this paper. First, we need to prevent the model from overfitting to a specific KG, i.e., the model's parameters are not much influenced by the content of an individual KG, while the common structural information of KGs can be well captured. To this end, we regard concept embeddings as additional features to enrich instances and use the prototypical network [11] to conduct instance matching. Since the entity-level information should not be ignored, we also take the entity descriptive information as another kind of features. Second, to capture the semantics of relation types for implicit relation matching, the model needs to rapidly summarize relation-specific information from limited instances with the help of concept-level KGs. Borrowing the idea in MetaR [2], we design a relation-meta learner to get the global relation representation given some head-tail entity pairs. The triplet loss here, together with the cross-entropy loss in the prototype network, forms the total loss of the training phase, and the corresponding scoring functions are combined in a similar way to calculate the distribution of relation types in the inference phase.

We adopt *WikiData* KG for the general domain [30] and *UMLS* KG [21] for the medical domain in this work, and evaluate KEFDA on the *FewRel 2.0* DA challenge [6]. Experimental results show that our approach achieves a leading performance over the state-of-the-art methods with significant improvements (6.63% on average).

In summary, the main contributions of this paper includes:

- It puts forward a **K**nowledge-**E**nhanced **F**ew-shot RC model for the **D**omain **A**daptation task, named KEFDA. To the best of our knowledge, this is the first work to incorporate domain-specific knowledge to the few-shot RC task without processing large-scale graph data.

[2]https://www.nlm.nih.gov/research/umls/index.html
[3]https://www.researchspace.org/index.html
[4]https://www.acemap.info/
[5]https://github.com/alicogintel/AliCoCo

Table 1: Examples of 2-way 1-shot RC tasks from the *FewRel 2.0* dataset. For each sentence, the head entity (in bold blue) and the tail entity (in italic *red*) are annotated in advance. Instances in the training and testing tasks come from different domains.

Training Task (Collected from Wikipedia)		
Support Set	(A) crosses	The **DeSoto Bridge** across the *Mississippi River*.
	(B) part_of	**Herm** is one of the *Channel Islands* in the English Channel.
Query	Jingkou District is one of three districts of **Zhenjiang** , *Jiangsu* province, China.	
Testing task (Collected from Biomedical Literature)		
Support Set	(A) classified_as	These **tumors** are the most common *non-epithelial neoplasms* of gastric wall.
	(B) occurs_in	Aniridia is a rare **congenital** ocular disorder of complete or partial *iris hypoplasia*.
Query	The lateral lesions and **dental cysts**, especially *radicular cysts*, are compared.	

- It proposes to use the concept-level KG to help better understand the semantics of relation types with only a few instances, and innovatively views the manner of using KGs as a kind of meta-information that can be transferred from the general domain to specific domains.
- KEFDA significantly outperforms the state-of-the-art models and all the participants on the Domain Adaptation Challenge of the *FewRel 2.0* benchmark[6].

In the following, an overview of few-shot RC, knowledge-enhanced models in NLP, and KG embedding is given in Section 2. Then, we introduce our model in detail in Section 3, and the effectiveness of KEFDA is demonstrated in Section 4. The paper is concluded with future work in Section 5.

2 RELATED WORK

2.1 Few-Shot Relation Classification

FewRel [9] is a large-scale few-shot relation classification benchmark, on which comprehensive evaluations of several vanilla few-shot learning methods have been conducted, such as *Meta Network* [24], *GNN* [28], *SNAIL* [23], and *Prototypical Network* [11]. Afterward, many improved approaches have been explored, such as transfer learning based methods [5, 34], pre-trained language model based methods [29], meta-learning methods [7, 26], and prototype-based methods [22, 37]. Notice that all of these models are trained and tested on datasets of the same domain, which requires sufficient labeled training tasks in each target domain to achieve satisfactory classification performance. However, as mentioned before, domain-specific training data are quite scarce in real scenarios.

To tackle this, *FewRel 2.0* [6] was presented, including the DA challenge, where the model is trained in the general domain while performs inference in the medical domain. We evaluate our model on this challenge, and find that there has been no published method to solve it yet so far.

2.2 Knowledge-Enhanced NLP Model

Because of rich semantic information, KGs have been widely employed in increasing works of NLP models during the pre-training phase [19, 20, 31]. For example, Peters et al. [27] proposed the use of external entity features and attention mechanism to enhance the representation of each word piece in a sentence. K-BERT [18] injects triples into sentences as domain knowledge for language representation learning. KEPLER [32] regards textual descriptions of entities as a bridge to jointly optimize the language representation model and the knowledge embedding model. ERNIE [40] fuses knowledge information into language representation by a knowledgeable aggregator via pre-training tasks.

However, since these approaches use large-scale entity-level KGs of general domain in the pre-training phase, they are not suitable for the DA task due to the necessity of pre-training again and the problem of large space occupation and heavy computational efforts. To alleviate this, we propose a lightweight approach, which uses domain-specific concept-level KG directly in the downstream task.

2.3 Knowledge Graph Embedding

Knowledge Graph Embedding (KGE) maps the elements of KGs into a low-dimensional vector space, which enables the rich semantic information in entities and relations to be used in knowledge-aware applications [13]. Here, a distance-based or similarity-based scoring function is employed to measure the plausibility of facts so that positive triples will get higher scores than negative ones [1, 17, 25, 36]. In particular, JOIE [10] proposes a two-view KG embedding model from both the ontology view for abstract commonsense concepts and the instance view for specific entities. We follow this idea and mainly utilize the ontology part of KGs.

It is worth noting that in recent years, there have been researches on relation representation learning in few-shot scenarios. For example, MetaR [2] was proposed for the few-shot link prediction in KG, which introduces **relation meta** to aggregate the relation-specific information from the factual triples of the support set and apply it to the query set. Inspired by this, we manage to summarize the support instances of a relation into a global relation meta with the help of a concept-level KG, in order to capture the implicit semantic information of relation types in a specific domain.

3 MODEL

In this section, we present a detailed description of the proposed model KEFDA, including notations, overall framework, and the two main components.

[6] https://competitions.codalab.org/competitions/27981#results

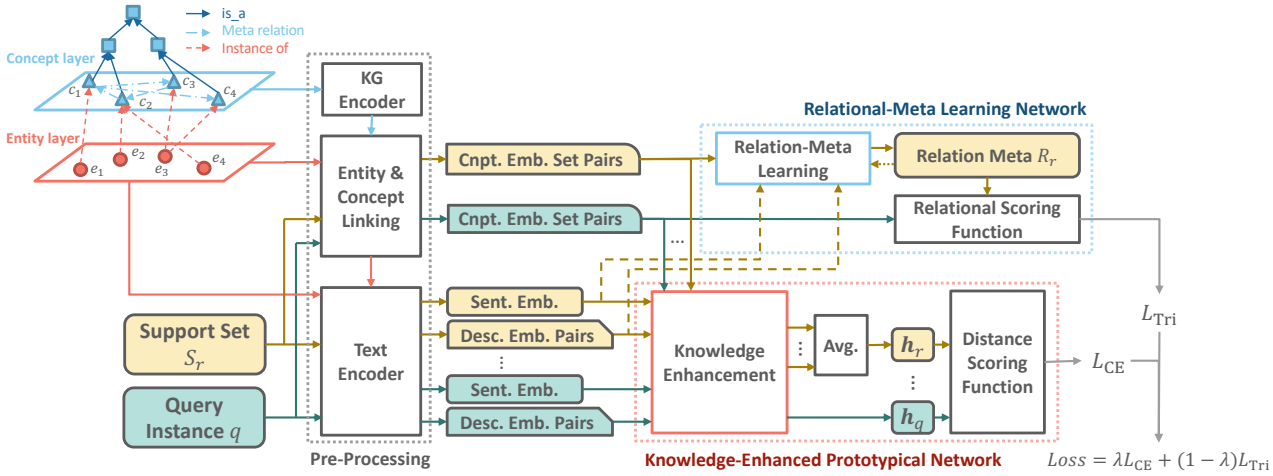

Figure 2: Framework of the proposed model KEFDA.

3.1 Notations

At first, we provide a formal definition of the knowledge graph and the few-shot relation classification task.

Definition 3.1 (Knowledge Graph). A KG can be written as $\mathcal{G} =< \mathcal{E}_\mathcal{G}, \mathcal{R}_\mathcal{G}, \mathcal{TP} >$, where $\mathcal{E}_\mathcal{G}$ is the node set and $\mathcal{R}_\mathcal{G}$ is the relation set. The triple set $\mathcal{TP} = \{(h, r, t) \in \mathcal{E}_\mathcal{G} \times \mathcal{R}_\mathcal{G} \times \mathcal{E}_\mathcal{G}\}$ includes the facts recorded in the KG.

According to the structure of KG, such as *WikiData* and *UMLS*, $\mathcal{G}$ can be divided into two subgraphs, namely the entity-level KG and the concept-level KG, as shown in Figure 1. Each entity node e_i can link to a set of concept nodes C_{e_i} (abbreviated as C_i). For instance, the entity `ADORA2b` is an instance of both `Protein` and `Receptor`. In addition to fact knowledge, each node in the KG has an aligned entity description as attribute, as shown under the node.

Definition 3.2 (Few-Shot Relation Classification). For an individual few-shot RC task $\mathcal{T} =< \mathcal{R}, \mathcal{S}, q >$, the goal is to predict a relation label $r_q \in \mathcal{R}$ for the query instance q based on a support set $\mathcal{S}$ with the pre-defined relation set $\mathcal{R}$. In the N-way K-shot setting, there are N relations, i.e., $\mathcal{R} = \{r_1, ..., r_N\}$, and for each $r \in \mathcal{R}$, the support subset $\mathcal{S}_r \subseteq \mathcal{S}$ contains K instances $\mathcal{S}_r = \{s_r^1, ..., s_r^K\}$. Each instance in the support set, as well as the query instance, consists of a sentence x and a head-tail entity pair (e_1, e_2).

Here, both N and K are usually quite small, so the support set of an individual task itself is not qualified as training data. Therefore, the training process of a few-shot RC model is based on a set of training tasks $\mathcal{T}_{\text{train}} = \{\mathcal{T}_r\}$, while its testing phase is conducted on a set of testing tasks $\mathcal{T}_{\text{test}} = \{\mathcal{T}_{r'}\}$. The relations occurring in the two kinds of tasks may be different, even in different domains. Table 1 shows 2-way 1-shot examples of few-shot RC tasks.

3.2 Framework

Essentially, for a few-shot RC task, the relation label of a query instance can be determined by computing the matching degree between the query instance and each relation in the support set. Our model KEFDA computes this degree from two perspectives: instance matching and implicit relation matching. **Instance matching** focuses on the contextual semantics of sentences and compare directly among instances, while **implicit relation matching** computes a relational meta for each relation type, which summarizes the global semantics of multiple instances with the help of a concept-level KG and use it to score the query instance. After that, these two matching degrees are combined by weighted sum, and the weights reflect which perspective is more concerned.

As shown in Figure 2, the framework consists of three parts: pre-processing, knowledge-enhanced prototypical network, and relation-meta learning network. We take $N \times K$ support instances and a query instance as the input of the N-way K-shot task. In the pre-processing phase, the sentence of each instance is first fed into a text encoder and transformed to a low-dimensional contextual sentence representation. Meanwhile, we apply an entity and concept linking module for the head and tail entities in each instance to obtain their corresponding entity nodes, and further associate them to sets of concept nodes in KG. Then, we can easily acquire the textual description of each entity and use the text encoder again to get its embedding as additional domain-specific features. For the more important concept features, we encode all the nodes and relations in the concept KG into distributed representations in advance so that the concept embedding set corresponding to each entity can be conveniently retrieved after the concept linking is performed.

For the instance matching, we design a **knowledge-enhanced prototypical network**, which takes the sentence contextual features, entity description features, and concept features of each instance as input to get its enhanced representation. Then, the matching degree between instances can be calculated by a distance-based scoring function, and the cross-entropy loss can be obtained.

Besides, a **relation-meta learning network** is employed to compute the degree of implicit relation matching. Here, a relation-meta learner is designed to infer the semantics of each relation type from pairs of concept embedding sets involved in the support set, utilizing embeddings of support sentences and concept descriptions. Then, a scoring function is used to measure the plausibility that a relation type exists between the head and tail entities in the query

instance, taking the corresponding pair of concept embedding sets as input. Similar to MetaR, the gradient meta is adopted to update the relation meta rapidly, and the triplet loss is used to update the parameters of the whole network.

Finally, we combine these two losses with a hyper-parameter to obtain the overall loss. Next, we will elaborate the two networks.

3.3 Knowledge-Enhanced Prototypical Network

Knowledge-enhanced prototypical network is designed for the calculation of instance matching degree, where the explicit contextual features of each instance are enhanced with domain-specific knowledge, i.e., the entity descriptions and concept-level relation information. This network consists of two components: the knowledge enhancement module and the prototypical network.

3.3.1 ***Knowledge Enhancement Module.*** For each instance $s = (x, e_1, e_2)$, given its contextual sentence representation $\mathbf{h}_x$ as well as its head-tail entity description features $\mathbf{h}^{\text{des}}_{e_1}$ and $\mathbf{h}^{\text{des}}_{e_2}$, we concatenate and put them into a feed forward neural network (FFNN),

$$\begin{aligned} \mathbf{h}^0 &= \mathbf{h}_x \oplus \mathbf{h}^{\text{des}}_{e_1} \oplus \mathbf{h}^{\text{des}}_{e_2}, \\ \mathbf{h}^l &= \sigma(\mathbf{W}^l \mathbf{h}^{l-1} + b^l), \\ \mathbf{h}' &= \mathbf{W}^L \mathbf{h}^{L-1} + b^L, \end{aligned} \tag{1}$$

where L is the number of layers in the network. $\mathbf{W}^l$ and b^l indicate the weights and bias of the l-th layer, $\mathbf{a} \oplus \mathbf{b}$ represents concatenation operation, and ReLU is used for activation function $\sigma(\cdot)$.

After that, since we have retrieved the concept set $C_i (i = 1, 2)$ for each entity e_i through concept linking, we compute the concept features $\mathbf{h}_{C_i}$ via an average operation on the concept embeddings $\mathbf{h}_c (c \in C_i)$, and also use FFNN to further fuse these features to get the final enhanced sentence representation $\mathbf{h}''$,

$$\begin{aligned} \mathbf{h}_{C_i} &= \frac{1}{|C_i|} \sum_{c \in C_i} \mathbf{h}_c, (i = 1, 2) \\ \mathbf{h}'' &= \text{FFNN}(\mathbf{h}' \oplus \mathbf{h}_{C_1} \oplus \mathbf{h}_{C_2}). \end{aligned} \tag{2}$$

For simplicity, we denote the whole enhancement process as:

$$\mathbf{h}_s = \mathbf{h}'' = f_\phi(s), \tag{3}$$

where ϕ includes the learnable parameters in the network above.

3.3.2 ***Prototypical Network.*** Prototypical network, a widely used solution for FSL, computes a representative prototype for each relation type of the support set. A simple and efficient way to calculate the prototype $\mathbf{h}_r$ of relation r is to average all the instances of the relation in the support set $\mathcal{S}_r$,

$$\mathbf{h}_r = \frac{1}{|\mathcal{S}_r|} \sum_{s \in \mathcal{S}_r} \mathbf{h}_s. \tag{4}$$

Then, the probability that the query instance q belongs to relation r can be calculated based on the distance between the query instance embedding $\mathbf{h}_q = f_\phi(q)$ and the prototype $\mathbf{h}_r$,

$$p_\phi(y = r|q) = \frac{\exp(-\text{d}(\mathbf{h}_q, \mathbf{h}_r))}{\sum_{r' \in \mathcal{R}} \exp(-\text{d}(\mathbf{h}_q, \mathbf{h}_{r'})}. \tag{5}$$

Since squared Euclidean distance can greatly improve results for prototypical networks [11], we adopt it as the distance metric $d(\cdot, \cdot)$ and its negation as the scoring function here as well. Finally, we use the cross-entropy loss L_{CE} for the network,

$$L_{\text{CE}} = -\sum_{(\mathcal{R}, \mathcal{S}, q) \in \mathcal{T}_{\text{train}}} \sum_{r \in \mathcal{R}} I_r \log p_\phi(y = r|q), \tag{6}$$

where I_r indicates whether the relation r is the ground-truth classification result. If so, $I_r = 1$; otherwise, $I_r = 0$.

3.4 Relation-Meta Learning Network

Besides capturing semantic information from the instances themselves, we believe that the domain-specific KG can also provide the semantics of relation types at the concept level as another clue for RC. That is realized by the relation-meta learning network, where the semantics of head-tail entity pairs in the support set is summarized to a relation meta with the help of concept-level KG, to determine whether the entity pair of query instance also embodies a similar implicit relation. In addition, a relation-meta updater is employed to rapidly modify the learned representation with support instances.

3.4.1 ***Relation-Meta Learner.*** In the task $\mathcal{T} = (\mathcal{R}, \mathcal{S}, q)$, for a support instance with head-tail entity pair $s = (x, e_1, e_2) \in \mathcal{S}_r \subseteq \mathcal{S}$, we pair the concepts in C_1 and C_2 into a collection $\mathcal{P}_s = \{(c_1^j, c_2^j) | c_1^j \in C_1 \wedge c_2^j \in C_2\}$, where (c_1^j, c_2^j) denotes the j-th head-tail concept pair for the instance. Then, FFNN is adopted again to get the concept-pair-specific relation meta $\mathbf{R}_j$,

$$\mathbf{R}_j = \text{FFNN}(\mathbf{h}_{c_1^j} \oplus \mathbf{h}_{c_2^j}). \tag{7}$$

Since each pair can derive such a concept-pair-specific relation meta, we need to aggregate them to get the instance-specific relation meta $\mathbf{R}_s$. It is worth noting that these concept pairs are not equally important for the current support instance, so we use the **attention mechanism** to emphasize the concept pair that is more relevant to the current instance.

Specifically, we use the description features of both head concept $\mathbf{h}^{\text{des}}_{c_1^j}$ and tail concept $\mathbf{h}^{\text{des}}_{c_2^j}$ as the candidate keys, the corresponding relation meta $\mathbf{R}_j$ as the value, and the contextual feature of support instance $\mathbf{h}_x$ as the query to make the attention-based aggregation among concept pairs. The higher score computed by the head or tail entity will be used as the attention weight of the concept-pair-specific relation meta. This process can be written as:

$$\begin{aligned} \mathbf{R}_s &= \sum_{(c_1^j, c_2^j) \in \mathcal{P}_s} p_j \mathbf{R}_j \\ p_j &= \text{softmax}(\alpha_j), \\ \alpha_j &= -\max(-\text{d}(\mathbf{h}_x, \mathbf{h}^{\text{des}}_{c_1^j}), -\text{d}(\mathbf{h}_x, \mathbf{h}^{\text{des}}_{c_2^j})). \end{aligned} \tag{8}$$

After that, we average the relation meta corresponding to each instance in the support set of a relation r, to obtain the final relation-specific relation meta $\mathbf{R}_r$ for the current task,

$$\mathbf{R}_r = \frac{1}{|\mathcal{S}_r|} \sum_{s \in \mathcal{S}_r} \mathbf{R}_s. \tag{9}$$

3.4.2 ***Relation-Meta Updater.*** Drawing on the idea of MetaR, we use gradient meta to make rapid corrections to the learned relation meta based on the support set of the task. Applying the key idea of similarity-based KGE, the bilinear scoring function for each concept pair $(c_1^j, c_2^j) \in \mathcal{P}_s$ of an instance $s \in \mathcal{S}_r$ is adopted:

$$g_r(c_1^j, c_2^j) = \mathbf{h}_{c_1^j}^\top \mathbf{R}_r \mathbf{h}_{c_2^j}, \tag{10}$$

Although some concept pairs have low relevance to the current support set, we find that around 90.13% of entities link to a unique concept node, which should be regarded valid. Therefore, we consider all the concept pairs as positive samples and use the triplet loss function to calculate the loss for $\mathcal{S}_r$,

$$L(\mathcal{S}_r) = \sum_{s \in \mathcal{S}_r} \sum_{(c_1^j, c_2^j) \in \mathcal{P}_s} \max(0, \gamma - g_r(c_1^j, c_2^j) + g_r(c_1^j, c_2')), \tag{11}$$

where the margin γ is a hyper-parameter, and (c_1^j, c_2') denotes a negative sample, whose tail concept c_2' is randomly replaced by another concept node and satisfies $(c_1^j, c_2') \notin \mathcal{P}_s$ for any $s \in \mathcal{S}_r$.

Following the gradient update rule, we rapidly update the relation meta $\mathbf{R}_r$ as follows:

$$\begin{aligned} G_r &= \nabla_{\mathbf{R}_r} L(\mathcal{S}_r), \\ \mathbf{R}_r' &= \mathbf{R}_r - \beta G_r, \end{aligned} \tag{12}$$

where G_r is the gradient meta of $\mathbf{R}_r$ and β indicates its step size. The updated relation meta $\mathbf{R}_r'$ is used to score the query instance $q = (x_q, e_1^q, e_2^q)$. Similar to the support set, the concept pairs of the query instance can be constructed, i.e., $\mathcal{P}_q = \{(c_1^q, c_2^q) | c_1^q \in C_1^q \wedge c_2^q \in C_2^q\}$. Then, the scoring function and the loss function can be written as:

$$\begin{aligned} g_r(c_1^q, c_2^q) &= \mathbf{h}_{c_1^q}^\top \mathbf{R}_r' \mathbf{h}_{c_2^q}, \\ L(q) &= \sum_{(c_1^q, c_2^q) \in \mathcal{P}_q} \max(0, \gamma - g_r(c_1^q, c_2^q) + g_r(c_1^q, c_2')). \end{aligned} \tag{13}$$

The total loss L_{Tri} of the relation-meta learning network is the sum of query losses calculated from each task in $\mathcal{T}_{\text{train}}$,

$$L_{\text{Tri}} = \sum_{(\mathcal{R}, \mathcal{S}, q) \in \mathcal{T}_{\text{train}}} L(q). \tag{14}$$

3.5 Training Objective

The final loss is a trade-off between the cross-entropy loss L_{CE} of knowledge-enhanced prototypical network, which reflects the instance matching degree between the prototype of the support set and the query instance, and the triplet loss L_{Tri} of relation-meta learning network, which indicates an implicit relation matching degree based on the concept-level KG.

$$Loss = \lambda L_{\text{CE}} + (1 - \lambda) L_{\text{Tri}}, \quad 0 \le \lambda \le 1, \tag{15}$$

where λ is the hyper-parameter to adjust the importance of the two perspectives of matching.

Table 2: Statistics of *FewRel 2.0* DA benchmark. *#Rel.*, *#Ins.*, and *#Ent.* denote the number of relation types, instances and entities respectively. *Hav. Des.* and *Hav. Cpt.* means the proportion of entities with textual descriptions and those with concept links to the corresponding KG respectively.

Task	#Rel.	#Ins.	#Ent.	Hav. Des.	Hav. Cpt.
Training	64	44, 800	89, 600	93.47 %	99.54 %
Validation	10	1, 000	2, 000	90.25 %	98.70 %
Testing	15	1, 500	3, 000	88.48 %	99.01 %

Table 3: Statistics of concept-level KGs we use. *#Node*, *#Rel.*, and *#Triple* denote the number of nodes, relation types, and triples included in the KG respectively.

KG	#Node	#Rel.	#Triple
WikiData (Ontology)	6409	39	8057
UMLS (Semantic Network)	127	54	5890

4 EXPERIMENTS

4.1 Benchmark Dataset & Knowledge Graph

We evaluate our model on the *FewRel 2.0* DA challenge[7] [6], whose training tasks in the general domain are the same as those in *FewRel*, while the validation and testing tasks were collected from the medical domain. Table 2 shows the statistics of this dataset, and the relation types that occur in the training, validation, and testing tasks are exclusive. Besides, the challenge maintains evaluation websites[8,9] for researchers to upload the prediction results of their models for comparison. For fairness concern, only the labels of training and validation tasks are publicly available in the dataset, so the ablation study and the baselines implemented by ourselves are performed upon the validation tasks.

We construct the general and domain-specific KGs based on *WikiData*[10] and *UMLS*[11] knowledge bases (KBs), respectively. As a free KB hosted by *Wikimedia*[12] and edited by volunteers, *WikiData* contains abundant structured and commonsense information. The *UMLS* KB, summarized and compiled by experts, includes detailed medical entity definitions, related concepts, and semantic relations among them, making it an ideal and easily accessed source of knowledge in the medical field. In addition, the both KBs own abundant textual descriptions of entities and concepts, which are convenient to build KGs meeting the requirement of our model.

Similar to JOIE [10], the both KGs are divided into an entity layer and a concept layer. Since only the concept-level KG is used for knowledge embedding, we do not need to download the complete KG as other works did, which is large and unsuitable to access online. Instead, we utilize *WikiData*'s SPARQL query service[13] to

[7] https://github.com/thunlp/FewRel
[8] https://thunlp.github.io/2/fewrel2_da.html
[9] https://competitions.codalab.org/competitions/27981
[10] https://www.wikidata.org/
[11] https://www.nlm.nih.gov/research/umls/index.html
[12] https://www.wikimedia.org/
[13] https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service

Table 4: Accuracy (%) of models on *FewRel 2.0* testing tasks under N-way K-shot settings, and the best results are in bold.

Few-Shot RC Model	Avg.	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
Proto (CNN)	35.67	35.09	49.37	22.98	35.22
Proto (BERT)	38.75	40.12	51.50	26.45	36.93
Proto-ADV (BERT)	40.35	41.90	54.74	27.36	37.40
Proto-ADV (CNN)	43.54	42.21	58.71	28.91	44.35
BERT-PAIR	66.93	67.41	78.57	54.89	66.85
PAMN	78.98	77.54	90.40	65.98	82.03
DualGraph	81.83	80.11	91.01	73.89	82.34
GTP	82.18	80.04	92.58	69.25	86.88
KEFDA	**88.82**	**87.81**	**95.00**	**81.84**	**90.63**

Table 5: Accuracy (%) of ERNIE and the variants of our model KEFDA on *FewRel 2.0* validation tasks under N-way K-shot settings. *(-Desc.), (-Cnpt.),* and *(-Meta.)* denote the absence of description features, concept features, and the relation-meta learning network, respectively.

Few-Shot RC Model	Avg.	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
ERNIE	54.26	55.24	62.70	47.68	51.43
KEFDA-DistMult (-Desc. -Cnpt. -Meta.)	53.24	58.63	63.08	33.64	57.60
KEFDA-DistMult (-Cnpt. -Meta.)	66.53	72.95	68.58	59.59	64.98
KEFDA-DistMult (-Meta.)	87.52	85.55	93.75	**80.38**	90.40
KEFDA-RotatE	64.69	60.82	76.92	50.82	70.19
KEFDA-TransE	67.48	62.82	80.98	53.69	72.43
KEFDA-ANALOGY	86.85	85.58	94.30	78.84	88.69
KEFDA-DistMult	**87.69**	**86.18**	**94.38**	79.46	**90.77**

construct the concept-level KG in general domain, and adopt the semantic network in *UMLS*[14] directly as the medical concept-level KG. Table 3 shows the statistics of the two concept-level KGs.

The entities appearing in *FewRel 2.0* were just annotated based on *WikiData* and *UMLS*, so it can be guaranteed that every entity in the dataset can be linked to the corresponding entity node in KG by a unified identifier. Therefore, we can easily retrieve its textual description and map it to a set of concept nodes online. Table 2 also presents the proportion of head and tail entities in the dataset for which there exist textual descriptions and the corresponding concepts in KG. We observe that the proportions are high enough to make our approach feasible, and about 99% of entities can be augmented with concept-level information, which creates favorable pre-conditions for our concept-based enhancement approach.

4.2 Experiment Settings

Following the dataset configuration, we consider all four types of few-shot settings given in the DA challenge: 5-way 1-shot, 5-way 5-shot, 10-way 1-shot, and 10-way 5-shot.

The text encoder we use is the pre-trained $BERT_{BASE}$ model[15], where the dimension of text embedding is 768, and the maximum length of the sentence is set to 128. We use DistMult as the KG encoder, which is implemented by the OpenKE toolkit[16] [35], and the dimension of the graph embedding is set to 256. When a textual description or concept cannot be found in KG, we pad it with an all-zero vector as substitute. Adam optimizer [15] is used with the initial learning rate 10^{-4}, and following the setting of MetaR, we set both the margin γ of triplet loss and the step size β of gradient meta to 1. After hyper-parameter analysis, we set the default value of λ to 0.95, and the detailed analysis is presented in Section 4.6. The model is trained for 30,000 epochs, and evaluated upon validation tasks every 1000 epochs. Among these saved models with current parameters, the one performing best on the validation tasks will be treated as the final model and used to conduct evaluations upon testing tasks. The source code and experiment details of KEFDA can be obtained from https://github.com/imJiawen/KEFDA.

4.3 Baselines

To verify the effectiveness of our approach, we select some representative methods on the few-shot RC task as baselines. The prediction results of these models were all uploaded and publicly exhibited on the evaluation website. The classical prototypical network is denoted as **Proto**, and there are many variants based on it. **(CNN)** and **(BERT)** represent the use of CNN [38] and BERT as the text

[14] https://semanticnetwork.nlm.nih.gov/
[15] https://github.com/huggingface/transformers
[16] https://github.com/thunlp/OpenKE

encoder, respectively. The model with **"-ADV"** means the use of adversarial training [8], which is a commonly-used domain adaptation method and has been proved efficiently in finding domain-invariant features. We also consider **BERT-PAIR** [6], which has achieved high accuracy on *FewRel*. Besides, there are several anonymous works which gained promising results, including **PAMN**, **Dual-Graph**, and **GTP**. Although there are no published papers of their works yet, we present their results for reference.

In addition to the model devoted to the RC task, we also compare our model with the knowledge-enhanced language representation model, **ERNIE**, which fuses the entity information in KG into texts during the pre-training process. To accommodate it to the few-shot learning scenario, we adopt ERNIE as an encoder for the prototypical network, and evaluate it on the validation tasks. Since ERNIE was pre-trained with the general domain KG *WikiData*, when utilized to encode domain-specific texts, the best way is to pre-train the model again with a domain-specific KG and the related corpus. However, massive corpus and large-scale entity-level KGs in a specific domain are not always available, and re-training the model requires a large amount of computational efforts. As a compromise, we still link entities in the medical dataset to nodes in *WikiData* in advance and make a padding if the entity cannot be found. Although statistical results indicate that only around 40% of the entities in the medical dataset can be linked, this is the only reasonable and practical way we can use ERNIE upon domain-specific datasets.

4.4 Overall Results

Table 4 shows the accuracy of all the models on official testing tasks under the four few-shot settings, as well as the average of them. The results demonstrate that our model dramatically improves the classification accuracy of the DA task for all settings, and raises GTP, the best model so far except ours, by 6.63% on average. That indicates domain-specific KGs, especially the concept-level part, can indeed help achieve better classification performance than simply using contextual semantic information.

In addition, as shown in Table 5, ERNIE performs worse than our approach, as the coverage of the general KG on medical entities is low. This implies that enhancing models with only the general domain KG is insufficient for a professional area due to the huge differences among domains, and the usage of KG in the pre-training phase restricts its effectiveness for domain adaptation.

4.5 Ablation Study

This subsection explores the contributions of the relation-meta learning network and individual features we use in the prototypical network, as well as the choice of KG encoder. We also replace the attention mechanism in relation-meta learning network with the averaging operation, but since the results are related to the hyper-parameter λ, we present and analyze that in Section 4.6.

4.5.1 ***Effect of Knowledge Enhancement & Relation-Meta Learning.*** In Table 5, we show the accuracy of the model KEFDA without knowledge enhancement module and relation-meta learning network **(-Desc. -Cnpt. -Meta.)**, with only entity text description features **(-Cnpt. -Meta.)**, with both description features and concept features **(-Meta.)**, and the full model **KEFDA**. The results indicate the model's performance is promoted with the presence of each additional feature and the employment of relation meta, of which the concept features are most helpful. This suggests that concept-level KGs can provide abundant implicit structured and global knowledge, which helps the model make correct judgments.

4.5.2 ***Knowledge Graph Encoder.*** Since KG embedding plays a crucial role in the model, we also test the impact of the KG encoder on the classification results. Besides **DistMult** [36], **TransE** [1], **ANALOGY** [17] and **RotatE** [25] are chosen as candidates.

We observe that DistMult and ANALOGY significantly outperform other encoders. Such results are not surprising because they are capable to handle multi-relational edges in KGs, i.e., multiple types of relations can exist between a pair of entities. Besides, the accuracy of DistMult is slightly higher than that of ANALOGY. The reason is probably that ANALOGY has more parameters and thus prefers domain-specific features to common structural features, which hinders domain adaptation.

4.6 Hyper-Parameter Analysis

In this subsection, we discuss the effect of the hyper-parameter λ, i.e., a trade-off between instance matching degree and implicit relation matching degree. We perform a grid search to find the optimal hyper-parameter in each few-shot setting. Specifically, values are first fetched between 0.6 and 0.9 with the step size 0.1, and a more fine-grained division between 0.9 and 1.0 is then performed with the step size 0.025. Figure 3 shows the performance among different λ values under the four settings. As mentioned above, the simplified aggregation operation (i.e., average) on concept pairs is also considered for comparison.

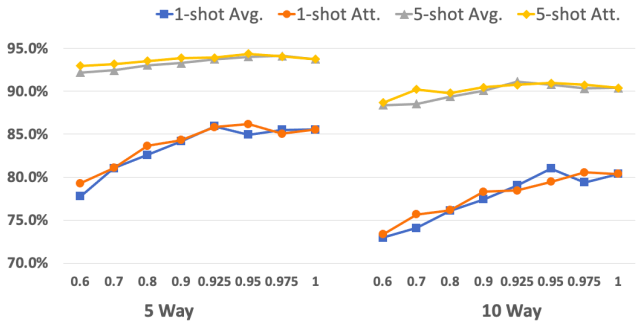

Figure 3: Accuracy on *FewRel 2.0* validation tasks under four settings with varied λ. *Att.* represents our KEFDA, and *Avg.* means a variant model replacing the attention mechanism in the relation-meta network with the average operation.

We can see that although the use of the attention mechanism does not always lead to better performance, it is more stable than the average operation when λ changes. Thus, we do not have to keep an eye on the selection of this hyper-parameter with the attention mechanism. Besides, it can be observed that the overall model achieves higher performance when λ locates within the interval $(0.9, 1.0)$, where the accuracy rises and then falls as λ increases. This illustrates that the implicit relational matching is indeed beneficial to the classification performance, but should not be over-concerned. Although the optimal value of λ is not always

the same under the four task settings, the differences are subtle, so it is reasonable to think that the model nearly exhibits its best performance when λ is set to 0.95. Moreover, we find that for 1-shot tasks, the classification performance drops dramatically with the decreasing of λ when λ is less than 0.9. This suggests when there are few support instances, the explicit semantic matching at the instance level is more important for classification, while attaching more focus to the implicit relation matching would degrade the performance. In contrast, as more instances are available (e.g., in 5-shot settings), the effect of implicit relation matching begins to manifest itself since the model can obtain more information about the global relation from the support set, and is thus able to compute a more exact relation meta with the help of concept-level KG.

5 CONCLUSION

This paper explores a new manner of utilizing prior knowledge to tackle the lack of training data in specific domains for the classical NLP task, relation classification from texts. This is the first attempt to leverage concept-level knowledge graph to capture global and implicit structural information of abstract concepts and relation types. This kind of meta information is able to be transferred across domains and can thus help solve few-shot machine learning tasks, especially in domain adaptation scenarios. The proposed model KEFDA takes up the first place until now on the Domain Adaptation Challenge of the *FewRel 2.0* benchmark. In the future, other domains besides medicine will be taken into account as long as a public knowledge graph of that domain can be easily accessed. We also plan to extend the model to more NLP tasks which require background knowledge of specific domains to complement insufficient labeled data.

ACKNOWLEDGMENTS

This work is supported by National Key R&D Program of China (2018YFC0116703), Strategic Priority Research Program of Chinese Academy of Sciences (XDC02060500), Youth Innovation Promotion Association CAS, and Zhejiang Lab (2020NF0AC02).

REFERENCES

[1] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NeurIPS*. 2787–2795.
[2] Mingyang Chen, Wen Zhang, Wei Zhang, Qiang Chen, and Huajun Chen. 2019. Meta Relational Learning for Few-Shot Link Prediction in Knowledge Graphs. In *EMNLP-IJCNLP*. 4216–4225.
[3] M. Cui, L. Li, Z. Wang, and M. You. 2017. A Survey on Relation Extraction. In *CCKS*. 50–58.
[4] Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement Learning for Relation Classification From Noisy Data. In *AAAI*. 5779–5786.
[5] T. Gao, X. Han, R. Xie, Z. Liu, F. Lin, L. Lin, and M. Sun. 2020. Neural Snowball for Few-Shot Relation Learning. In *AAAI*, Vol. 34. 7772–7779.
[6] T. Gao, X. Han, H. Zhu, Z. Liu, P. Li, M. Sun, and J. Zhou. 2019. FewRel 2.0: Towards More Challenging Few-Shot Relation Classification. In *EMNLP-IJCNLP*. 6249–6254.
[7] X. Geng, X. Chen, K. Q. Zhu, L. Shen, and Y. Zhao. 2020. MICK: A Meta-Learning Framework for Few-shot Relation Classification with Small Training Data. In *CIKM*. 415–424.
[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR*.
[9] X. Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, and M. Sun. 2018. FewRel: A Large-Scale Supervised Few-shot Relation Classification Dataset with State-of-the-Art Evaluation. In *EMNLP*. 4803–4809.
[10] J. Hao, M. Chen, W. Yu, Y. Sun, and W. Wang. 2019. Universal Representation Learning of Knowledge Bases by Jointly Embedding Instances and Ontological Concepts. In *SIGKDD*. 1709–1719.
[11] Snell Jake, Swersky Kevin, and Zemel Richard. 2017. Prototypical Networks for Few-shot Learning. In *NeurIPS*. 4077–4087.
[12] Emily Jamison. 2011. Using Grammar Rule Clusters for Semantic Relation Classification. In *RELMS@ACL*. 46–53.
[13] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2020. A Survey on Knowledge Graphs: Representation, Acquisition and Applications. *CoRR* abs/2002.00388 (2020).
[14] Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *ACL*. 178–181.
[15] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.
[16] J. Lee, S. Seo, and Y. Choi. 2019. Semantic Relation Classification via Bidirectional LSTM Networks with Entity-aware Attention using Latent Entity Typing. *Symmetry* 11 (2019), 785.
[17] Hanxiao Liu, Yuexin Wu, and Yiming Yang. 2017. Analogical Inference for Multi-Relational Embeddings. In *ICML*. 2168–2178.
[18] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang. 2020. K-BERT: Enabling Language Representation with Knowledge Graph. In *AAAI*. 2901–2908.
[19] Robert L. Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's Wife Hillary: Using Knowledge-Graphs for Fact-Aware Language Modeling. In *ACL*. 5962–5971.
[20] Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish A. Talati, and Ross W. Filice. 2019. Ontology-Aware Clinical Abstractive Summarization. In *SIGIR*. 1013–1016.
[21] Alexa T. McCray. 2003. An Upper-Level Ontology for the Biomedical Domain. *Comparative and Functional Genomics* 4 (2003), 80–84.
[22] Fan Miao, Bai Yeqi, Sun Mingming, and Li Ping. 2019. Large Margin Prototypical Network for Few-shot Relation Classification with Fine-grained Features. In *ACM*. 2353–2356.
[23] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A Simple Neural Attentive Meta-Learner. In *ICLR*.
[24] Tsendsuren Munkhdalai and Hong Yu. 2017. Meta Networks. In *ICML*, Vol. 70. 2554–2563.
[25] Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. 2016. Holographic Embeddings of Knowledge Graphs. In *AAAI*. 1955–1961.
[26] Abiola Obamuyide and Andreas Vlachos. 2019. Model-Agnostic Meta-Learning for Relation Classification with Limited Supervision. In *ACL*. 5873–5879.
[27] Matthew E. Peters, Mark Neumann, Robert L. Logan, Roy Schwartz, V. Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *EMNLP-IJCNLP*. 43–54.
[28] Victor Garcia Satorras and Joan Bruna. 2018. Few-Shot Learning with Graph Neural Networks. In *ICLR*.
[29] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *ACL*. 2895–2905.
[30] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57 (2014), 78–85.
[31] Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. 2017. Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification. In *IJCAI*. 2915–2921.
[32] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics* 9 (2021), 176–194.
[33] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. 2020. Generalizing from a Few Examples: A Survey on Few-Shot Learning. *Comput. Surveys* 53 (2020), 63:1–63:34.
[34] Ruidong Wu, Yuan Yao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2019. Open Relation Extraction: Relational Knowledge Transfer from Supervised Data to Unsupervised Data. In *EMNLP-IJCNLP*. 219–228.
[35] Han Xu, Cao Shulin, Lv Xin, Lin Yankai, Liu Zhiyuan, Sun Maosong, and Li Juanzi. 2018. OpenKE: An Open Toolkit for Knowledge Embedding. In *EMNLP*. 139–144.
[36] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR*.
[37] Kaijia Yang, Nantao Zheng, Xinyu Dai, Liang He, Shujian Huang, and Jiajun Chen. 2020. Enhance Prototypical Network with Text Descriptions for Few-shot Relation Classification. In *CIKM*. 2273–2276.
[38] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao. 2014. Relation Classification via Convolutional Deep Neural Network. In *COLING*. 2335–2344.
[39] Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional Long Short-Term Memory Networks for Relation Classification. In *PACLIC*. 73–78.
[40] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *ACL*. 1441–1451.
[41] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *ACL*. 207–212.