小样本关系分类研究综述

胡晗　　刘鹏远*
北京语言大学　信息科学学院
国家语言资源监测与研究平面媒体中心
北京市海淀区学院路15号, 100083
201821198609@stu.blcu.edu.cn　liupengyuan@blcu.edu.cn

摘要

关系分类作为构建结构化知识的重要一环，在自然语言处理领域备受关注。但在很多应用领域中(如医疗、金融等领域)，收集充足的用于训练关系分类模型的数据十分困难。近年来，仅需要少量训练样本的小样本学习逐渐应用于关系分类研究中。本文对近期小样本关系分类模型与方法进行了系统的综述。根据度量方法的不同，将现有方法分为原型式和分布式两大类。根据是否利用额外信息，将模型分为预训练和非预训练两大类。此外，除了常规设定下的小样本学习，本文还梳理了跨领域和稀缺资源场景下的小样本学习，探讨了目前小样本关系分类方法的局限性，并分析了跨领域小样本学习面临的技术挑战。最后，展望了小样本关系分类未来的发展方向。

关键词： 关系分类；小样本学习；元学习

Few-Shot Relation Classification: A Survey

Han Hu　　Pengyuan Liu*
Beijing Language and Culture University, School of Information Science
Language Resources Monitoring and Reserch Center
15 Xueyuan Road, Haidian District, Beijing, 100083, China
201821198609@stu.blcu.edu.cn　liupengyuan@blcu.edu.cn

Abstract

As an important part of constructing structured knowledge, relation classification has attracted much attention in the field of natural language processing. However, in many application fields (medical and financial fields), it is very difficult to collect sufficient data for training relation classification model. In recent years, few-shot learning research which only needs a small number of training samples is emerging in various fields. In this paper, the recent models and methods of few-shot relation classification are systematically reviewed. According to the different measurement methods, the existing methods are divided into prototype and distributed. According to whether to use additional information, the model is divided into two categories: pretraining and non-pretraining. In addition to the regular setting of few-shot learning, we also comb the cross domain few-shot learning and few-few-shot learning, and discusse the limitations of current few-shot relation classification methods, and analyze the technical challenges faced by cross domain few-shot models. Finally, the future development of few-shot relation classification is prospected.

Keywords: Relation Classification , Few-shot Learning , Meta Learning

* 通讯作者　Corresponding Author

第十九届中国计算语言学大会论文集，第363页-第375页，海南，中国，2020年10月30日至11月1日。

1 引言

关系分类是自然语言处理领域中的一项重要任务，它致力于判断给定语句中两个目标实体之间的预定义关系，为构建结构化知识(如，知识图谱)提供了基础。当前用于该任务的主流深度学习模型以大量监督数据为驱动，导致模型泛化能力依赖于监督数据的数量和质量。尽管正则技术被广泛用来降低深度学习模型对训练数据的过拟合，但其并不能为模型提供额外的监督信息。因此当监督数据不足时，简单地对模型加以正则并不能真正解决泛化问题(Wang et al., 2019b)。为了缓解训练数据不足的问题，节省人工标注成本，Mintz et al. (2009)采用了远程监督的方法。文章假设“两个实体如果在知识库中存在某种关系，则包含这两个实体的语句在某种程度上能表示出这种关系”，启发式地将语句中的目标实体与知识库中的实体对齐，达到自动标注语句的目的。但这个假设也带来了后续的问题：(1)同一实体对在不同语句中所蕴涵的关系可能不同，利用远程监督方法会产生噪声数据(如Figure 1所示)；(2)很多领域的知识库并不完善(如，医疗领域)，且大部分实体对和关系呈长尾分布，通过这种方法获取的可用于训练的数据仍然不足(如Figure 2所示)。

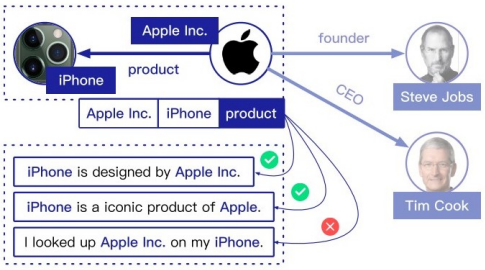

Figure 1: 远程监督方法引入了噪声数据。(Han et al. (2020))

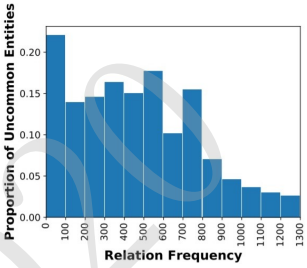

Figure 2: DBpedia中关系出现的频率与对应未见实体占比分布图。(Wang et al. (2019c))

相比之下，人类拥有利用过去所学知识快速学习新概念的能力。因此，研究者们希望构建一种新的训练方法，使模型仅在少量训练样本的情况下学习，并具备良好的泛化能力。Li Fei-Fei et al. (2006)首次提出单样本学习(One-Shot Learning)，采用贝叶斯模型，利用已学习的类别知识帮助模型在每个新类别仅有单个训练样本的情况下进行学习。至今，已有大量研究工作投入到单/小样本学习(One/Few-Shot Learning)领域，其中最具有代表性且主流的方法是元学习(Meta Learning)方法。元学习，或称“学会学习”，是系统地观察模型在不同的学习任务中的表现，从这种经验或元数据(Meta Data)中学习，然后以更快的速度学习新任务的方法(Vanschoren, 2018)。

目前，小样本学习的研究主要集中于计算机视觉领域。启发于人类的记忆，研究者们提出记忆网络，将先验知识存储在记忆模块中以供检索与更新(Weston et al., 2015; Sukhbaatar et al., 2015)。从优化的角度出发，一些研究者训练一个元优化器，帮助模型高效搜索合适的任务参数(Andrychowicz et al., 2016; Li and Malik, 2016)。另一些研究者则通过学习一个与任务无关的通用初始化参数，使得模型仅在少量训练样本情况下快速适应新任务(Finn et al., 2017)。Vinyals et al. (2016)从度量的角度提出了匹配网络，并首次提出了训练与测试过程相匹配的Episode训练原则(如Figure 3所示)。

在自然语言处理领域，小样本学习刚刚兴起。Yu et al. (2018)利用多个度量函数来解决任务多样性小样本分类问题。Geng et al. (2019)和Geng et al. (2020a)提出静态和动态记忆的归纳网络来解决因类别样本过少带来的样本方差问题。Han et al. (2018)首次将小样本学习引入关系分类任务，构建了小样本关系分类数据集FewRel，并尝试了几种典型的小样本学习方法与人类基准作比较。许多研究者在此基础上进行了探索，Baldini Soares et al. (2019)提出的无监督句子匹配方法在这一任务上的表现甚至超越了人类基准。针对小样本关系任务的多样性及任务中可能存在的噪声样本，Gao et al. (2019a)利用层级注意力来增强模型对小样本任务多样性以及噪声样本的鲁棒性。Xie et al. (2020)则通过异构图网络与对抗训练减少模型对噪声样本的敏感性。Obamuyide and Vlachos (2019)将监督式关系分类任务视为元学习的一个例

子，提出模型无关的元学习方案，力求模型在数据充足与数据稀缺两种情况下都有良好表现。由于一些领域的元数据不足以训练一个在该领域任务间有较好泛化能力的元模型，Gao et al. (2019b)在FewRel数据集的基础上提出了FewRel2.0数据集，探索元学习跨领域泛化以及非预定义类别检测问题。Geng et al. (2020b)则提出了更严苛的元训练条件，探索元学习模型在有限元数据情况下的学习能力。

在这篇综述文章中，我们系统地回顾了小样本关系分类任务具有代表性和启发性的工作(如Figure 4所示)。探讨了这些工作在当前用于解决该任务的元学习设定下的优势与不足。并给出了未来小样本关系分类的一些发展方向。

2 问题定义

2.1 N-way K-shot小样本分类

小样本学习是监督式机器学习的一种特殊情况，它的目标是在限制了目标任务训练数据数量的情况下，训练出对该任务新数据具有良好泛化能力的模型。

对于一个*N-way K-shot*小样本分类任务$T = \{D^{train}, D^{test}, \ell\}$，其中训练集(或称支持集)$D^{train} = \{(x_1^1, y_1), ..., (x_1^k, y_1), ..., (x_n^1, y_n), ..., (x_n^k, y_n)\}$ 包含N个类别(N一般为5或10)，每个类别K个训练样本(K一般为5或10)，测试集(或称问题集)$D^{test} = \{(x^m, y^m)\}_{m=1}^{M}$，$\ell$为损失函数。假设输入$x$和输出$y$的联合概率分布为$p(x, y)$，学习的最终目的是通过训练集$D^{train}$与损失函数$\ell$发现从$x$映射到$y$的最优假设$o^*$，且在测试集$D^{test}$上有良好的泛化能力。

由于训练集所提供的训练数据有限，用经验风险近似期望风险不够精准。当前以数据驱动为主的深度学习方法在这种任务上会出现过拟合的现象。尽管正则技术被广泛用来降低深度学习模型对训练数据的过拟合，但其并不能为模型提供额外的监督信息。因此，正则方法并不能提高小样本情况下用经验风险替代期望风险的可靠性(Wang et al., 2019b)。为了提高小样本情况下模型的泛化能力，结合先验知识至关重要。

2.2 元学习

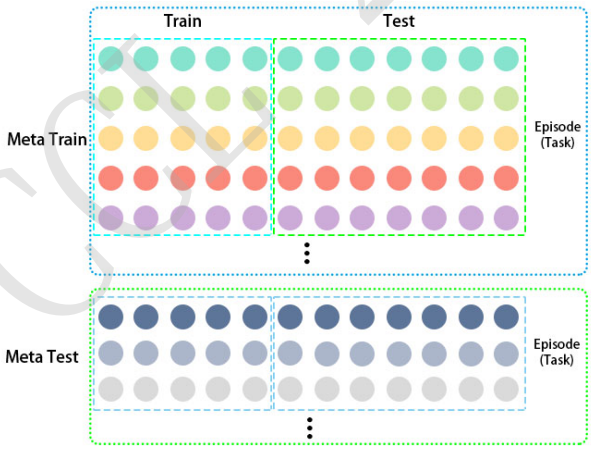

Figure 3: 元学习Episode训练框架

元学习，或称“学会学习”，是元学习器(Meta Learner)系统地观察基学习器(Base Learner)在不同的学习任务(Task)中的表现，从这种经验或元数据(Meta Data)中学习，然后以更快的速度学习未曾见过的新任务(Novel Task)的方法(Vanschoren, 2018)。这个过程中，存在两个层面的学习：(1)元学习器迭代地学习不同任务间的元知识(Meta Knowledge)；(2)基学习器基于元知识以及新任务中的特定信息快速学习并处理该任务。

对于一个小样本分类任务，基学习器的目标是找到最优假设o^*。为了接近o^*，基学习器确定了假设空间$\mathcal{H}$，其中包含了由φ参数化的假设$h(\cdot, \varphi)$。优化算法通过搜索假设空间$\mathcal{H}$来找到一个对于D^{train}最优的假设h。Wang et al. (2019b)系统地分析了经验风险的可靠性与样本复杂度和假设空间之间的联系。作者指出，为了使经验风险对期望风险的近似以一定概率达到一定精度，模型决定的假设空间越复杂，所需要的训练样本就越多。

在元学习中，为了减少训练所需的样本，元学习器需要从大量相似任务中学习元知识。然后，根据元知识构建假设空间$\mathcal{H}$的草图，以限制假设空间的大小。而基学习器则通过新任务特定的信息完成$\mathcal{H}$的具体构建。假设$p(T)$为小样本任务$T = \{D^{train}, D^{test}, \ell\}$的分布。在元训练阶段，元学习器$f_\theta(\cdot)$从包含$N_{meta}^{train}$个独立同分布任务的元训练集$D_{meta}^{train} = \{T_s^i \sim p(T)\}_{i=1}^{N_{meta}^{train}}$中学习。基学习器$g_\varphi(\cdot)$根据元知识从$D_{T_s}^{train}$中学习，并在$D_{T_s}^{test}$上评估损失。通过最小化基学习器在一系列任务上的损失来优化元学习器的参数θ:

$$\theta = \arg\min_\theta \mathbb{E}_{T_s \sim p(T)} \ell_\theta(D_{T_s}) \tag{1}$$

在元测试阶段，与元训练集类别互斥的元测试集$D_{meta}^{test} = \{T_t^j \sim p(T)\}_{j=1}^{N_{meta}^{test}}$被用来测试元学习器对新的小样本任务的泛化能力。

从监督式机器学习的角度，Chao et al. (2020)给出了更直观的解释。作者认为可以将元训练阶段视为元学习器根据一组$\{(D^{train}, h^*)\}$元样本对进行监督式训练的过程。给定一个小样本任务，假设基学习器在该任务上的最优假设为h^*，而基学习器根据元知识及训练集学习到的假设为h。给定该任务的测试集$D^{test} = \{(x^m, y^m)\}_{m=1}^M$，元学习器在该任务上的损失为$\ell_{meta}(g_\varphi(D^{train}|\theta), h^*) = |\mathcal{L}^{test}(h) - \mathcal{L}^{test}(h^*)|$，其中$\mathcal{L}^{test}(h) = \frac{1}{M}\sum_{m=1}^M \ell(h(x^m), y^m)$。假设$h^*$最小化$\mathcal{L}^{test}$，那么$\ell_{meta}(g_\varphi(D^{train}|\theta), h^*) = \mathcal{L}^{test}(h)$，即我们可以用$D^{test}$和$\ell$代替$h^*$与$\ell_{meta}$。因此，用于训练元学习模型的训练集由$\{(D^{train}, h^*)\}$变为了$\{(D^{train}, D^{test})\}$。最终，式(1)可以改写为:

$$\theta = \arg\min_\theta \sum_{i=1}^{N_{meta}^{train}} \sum_{m=1}^{M} \ell(g_\varphi(D_i^{train}|\theta)(x_i^m), y_i^m) \tag{2}$$

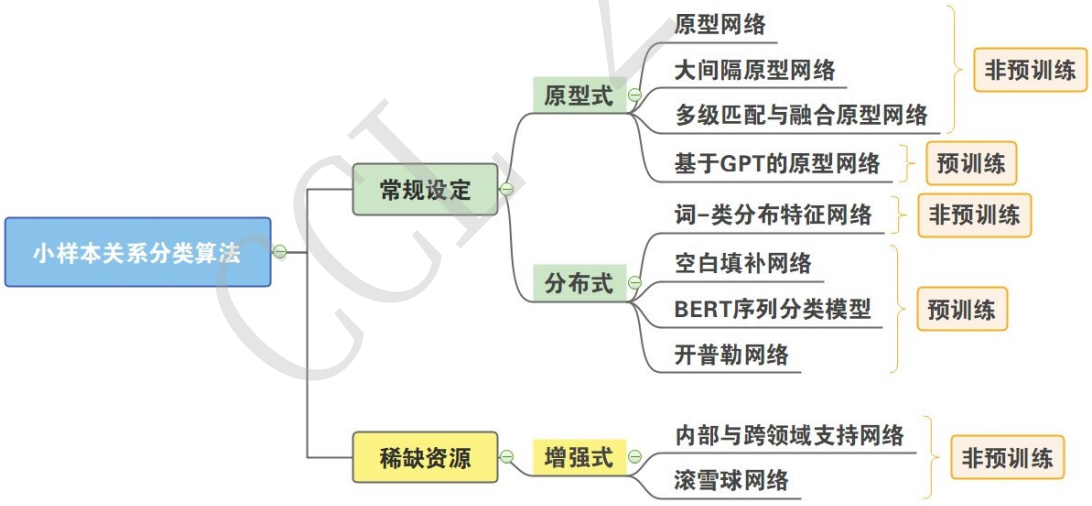

Figure 4: 小样本关系分类算法分类导图

3 常规小样本关系分类

3.1 数据集

FewRel(Han et al., 2018)是第一个英文小样本关系分类数据集，它包含100种关系，每种关系700个样本。作者以Wikipedia为数据库，Wikidata为知识库，通过远程监督的方法将数据库中的句子与知识库中的事实对齐。为了扩大实体集，作者首先利用命名实体识别技术挖掘文章中的非锚点实体，然后通过实体链接技术将挖掘出的实体与Wikidata中的实体进行匹配。由于对于表达某种关系的一组句子来说，其包含的可能是同一对实体。为了避免模型机械地根据句子中出现的实体对而不是句子本身的语义来进行关系分类，作者在每种关系中，对于同一对实体只保留一个样本。之后，去除样本量不足1000的关系，对剩余的关系，每种关系随机抽

取1000个样本。经过标注人员的筛选标注，去除正样本不足700的关系，以kappa值对剩余的关系进行降序排列，保留前100种关系。最终，数据集以64:16:20的比例被划分为训练集、验证集和测试集。

3.2 常规小样本关系分类算法

在常规小样本关系分类算法中，基于度量和优化的元学习方法最为常见。Han et al. (2018)测试了基于参数生成的元学习MetaNet(Munkhdalai and Yu, 2017)，基于图网络的元学习GNN(Satorras and Bruna, 2017)，基于时序卷积的元学习SNAIL(Mishra et al., 2017)。但这些复杂的方法在小样本关系分类任务上的表现并不如简单的基于度量的方法。后续的研究者在此基础上进行探索，我们将这些模型分为基于原型和基于分布式表达两大类。

3.2.1 原型式小样本关系分类算法

原型式小样本关系分类算法是基于度量的一类算法。度量方法将样本嵌入到一个更小的空间中，使得相似的样本聚在一起，不相似的样本分离。这些方法的不同点在于用于生成类别原型的向量表示以及生成类别原型的方法。

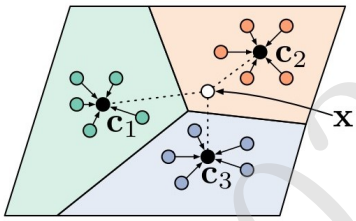

Figure 5: 原型网络(图片来自Snell et al. (2017))

- **原型网络(Prototypical Network(Snell et al., 2017))**假设存在一个嵌入空间，在这个空间里，每个类别的点都围绕着该类别的原型聚集(如Figure 5所示)。它利用卷积神经网络(CNN)作为编码器f_θ，将$x_n^k \in D^{train}$和$x^{test} \in D^{test}$非线性映射到该嵌入空间，然后构造其类别原型:

$$c_n = \frac{1}{K}\sum_{k=1}^{K} f_\theta(x_n^k) \tag{3}$$

  最后衡量$f_\theta(x^{test})$与c_n之间的距离$d(f_\theta(x^{test}), c_n)$（如，欧式距离）对其作最近邻分类。最终，通过分类损失对嵌入空间进行优化。

  在原型网络中，编码器$f_\theta(\cdot)$既是元学习器，也是基学习器。类似于多任务学习，它假设如果学习到的嵌入空间能够处理很多任务，那么这个空间也有足够的能力处理新任务。支持集不再用于基学习器的参数更新，而是作为嵌入空间中的类别锚点。

- **大间隔原型网络(Large Margin Prototypical Network(Fan et al., 2019))**在原型网络的基础上，采用了更加细粒度的特征表示以及额外的目标函数。除了利用句子级别的表示$f_{sentence}(x) = f_\theta(x)$以外，作者根据关系分类的特点，将句子分为五个部分，头实体之前的部分r_f，头实体e_h，头实体和尾实体之间的部分r_m，尾实体e_t，尾实体之后的部分r_b，利用多个CNN对其分别作嵌入得到嵌入表示$\mathbf{r}_f$，$\mathbf{e}_h$，$\mathbf{r}_m$，$\mathbf{e}_t$和$\mathbf{r}_b$。之后，将得到的表示拼接起来送入一个全连接层并用ReLU激活，获取这些表示的非线性关系:

$$f_{phrase}(x) = \mathbf{ReLU}(f_\varphi(\mathbf{r}_f \oplus \mathbf{e}_h \oplus \mathbf{r}_m \oplus \mathbf{e}_t \oplus \mathbf{r}_b)) \tag{4}$$

  然后将的句子级表示和短语级表示拼接起来得到最终的表示:

$$f(x) = f_{sentence}(x) \oplus f_{phrase}(x) \tag{5}$$

为了在嵌入空间中增加类间距离，缩短类内距离，作者额外采用了三元组损失函数作为辅助：

$$\mathcal{L}_{triplet} = \sum_{i=1}^{N} \mathbf{max}(0, \gamma + \|f(a_i) - f(p_i)\|^2) + \|f(a_i) - f(n_i)\|^2)) \tag{6}$$

其中，N为Episode/Task的大小，$a_i = c_n$是锚点，p_i是正样例，n_i是负样例。平衡交叉熵损失与三元组损失得到最终的损失函数：

$$\mathcal{L} = \mathcal{L}_{softmax} + \lambda \mathcal{L}_{triplet} \tag{7}$$

- **多级匹配与融合原型网络(MLMAN(Ye and Ling, 2019))**利用注意力机制，通过考虑支持集D^{train}和问题样本x^{test}的局部和实例两个层面的匹配信息，对两者作交互式编码。作者首先利用CNN对D^{train}和x^{test}进行上下文编码，得到$\{\mathbf{S}_k \in \mathbb{R}^{T_k \times d_c}; k = 1, ..., K\}$与$\mathbf{Q} \in \mathbb{R}^{T_q \times d_c}$。然后，将支持集拼接得到整个支持集的矩阵表示$\mathbf{C} \in \mathbb{R}^{T_s \times d_c}, T_s = \sum_{k=1}^{K} T_k$。

  对支持集与问题样例作局部匹配与融合：

$$\mathbf{A} = \mathbf{Q}\mathbf{C}^\top \in \mathbb{R}^{T_q \times T_s} \tag{8}$$

$$\widetilde{\mathbf{Q}} = \mathbf{Softmax}(\mathbf{A})\mathbf{C} \in \mathbb{R}^{T_q \times d_c} \tag{9}$$

$$\widetilde{\mathbf{C}} = \mathbf{Softmax}(\mathbf{A}^\top)\mathbf{Q} \in \mathbb{R}^{T_s \times d_c} \tag{10}$$

$$\bar{\mathbf{Q}} = \mathbf{ReLU}([\mathbf{Q}; \widetilde{\mathbf{Q}}; |\mathbf{Q} - \widetilde{\mathbf{Q}}|; \mathbf{Q} \odot \widetilde{\mathbf{Q}}]\mathbf{W}_1) \in \mathbb{R}^{T_q \times d_h} \tag{11}$$

$$\bar{\mathbf{C}} = \mathbf{ReLU}([\mathbf{C}; \widetilde{\mathbf{C}}; |\mathbf{C} - \widetilde{\mathbf{C}}|; \mathbf{C} \odot \widetilde{\mathbf{C}}]\mathbf{W}_1) \in \mathbb{R}^{T_s \times d_h} \tag{12}$$

将$\bar{\mathbf{C}}$还原为独立的类别矩阵$\{\bar{\mathbf{S}}_k \in \mathbb{R}^{T_k \times d_h}\}_{k=1}^{K}$，并用单层双向长短时记忆网络(BiLSTM)编码所有的$\bar{\mathbf{S}}_k$与$\bar{\mathbf{Q}}$，得到最终的局部表示$\hat{\mathbf{S}}_k$ 和$\hat{\mathbf{Q}}$。对局部表示进行最大池化和平均池化并拼接，得到最终向量表示：

$$\hat{\mathbf{s}}_k = [\mathbf{max}(\hat{\mathbf{S}}_k); \mathbf{ave}(\hat{\mathbf{S}}_k)], \forall k \in \{1, ...K\} \tag{13}$$

$$\hat{\mathbf{q}} = [\mathbf{max}(\hat{\mathbf{Q}}); \mathbf{ave}(\hat{\mathbf{Q}})] \tag{14}$$

除了对支持集与问题样例作局部匹配与融合，作者还采用了实例级匹配来构造类别原型。不同于原始原型网络通过平均向量的方式构造类别原型，MLMAN通过多层感知机度量$\hat{\mathbf{s}}_k$与$\hat{\mathbf{q}}$匹配程度，赋予不同$\hat{\mathbf{s}}_k$不同的权重来构造带权平均类别原型：

$$\beta_k = \mathbf{v}^\top(\mathbf{ReLU}(\mathbf{W}_2[\hat{\mathbf{s}}_k; \hat{\mathbf{q}}])) \tag{15}$$

$$\hat{\mathbf{s}} = \sum_{k=1}^{K} \frac{e^{\beta_k}}{\sum_{k'=1}^{K} e^{\beta_{k'}}} \hat{\mathbf{s}}_k \tag{16}$$

最终进行类别匹配，对问题样例作出分类：

$$f(\{s_k\}_{k=1}^{K}, q) = \mathbf{v}^\top(\mathbf{ReLU}(\mathbf{W}_2[\hat{\mathbf{s}}; \hat{\mathbf{q}}])) \tag{17}$$

为了生成更具代表性的类别原型，除了分类损失外，作者额外加入了非一致性度量损失，保证同一类别中的样本不会互相偏离：

$$\mathcal{L}_{incon} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \|\hat{\mathbf{s}}_k^i - \hat{\mathbf{s}}^i\|_2^2 \tag{18}$$

$$\mathcal{L} = \mathcal{L}_{softmax} + \lambda \mathcal{L}_{incon} \tag{19}$$

- **基于GPT的原型网络(Prototypical GP-Transformer(Eberts, 2019))**采用预训练语言模型GPT替代原始原型网络中的CNN作为编码器，以获得更好的类别原型的表示。在GPT中每个句子首尾有标记符标记句子的开始$< Start >$和结束$< End >$，由于Transformer是自注意力模型，$< end >$能够注意到整个句子，因此其嵌入表示$\mathbf{h}_{<end>}$被用于后续的分类。为了标示出句子中的目标实体，作者尝试了不同的标记目标实体的方法：(1)在目标实体两侧添加标记符(常用于RNN)；(2)位置嵌入(常用于CNN)；(3)将目标实体的平均嵌入表示与句子的平均嵌入表示拼接；(4)根据目标实体划分句子作分段编码并拼接；(5)将实体的平均嵌入表示与$\mathbf{h}_{<end>}$拼接。为了加速模型的收敛，作者在任务微调阶段加入了语言模型作为辅助目标函数：

$$\mathcal{L} = \mathcal{L}_{softmax} + \lambda \mathcal{L}_{LM} \tag{20}$$

3.2.2 分布式小样本关系分类算法

分布式小样本关系分类算法主要分为两类，一类是建模句子间的分布式表示，另一类是建模句子中词对类的分布式表示。

- **空白填补网络(Matching The Blanks(Baldini Soares et al., 2019))**将Harris分布式假设拓展到关系领域，利用预训练语言模型BERT，从无标注非结构化文本中学习任务无关的关系表示。作者假设，对于任意一对关系陈述句$\mathbf{r}$和$\mathbf{r}'$，如果它们表示的关系语义相似，那么两者的内积$f_\theta(\mathbf{r})^\top f_\theta(\mathbf{r}')$应该很大，否则很小。作者观察到，在网络文本中，任意一对实体之间的每种关系都可能被陈述多次。利用这一冗余特性，作者运用实体链接方法构建了无监督数据集，提出了名为Matching The Blanks的方法来学习判断两个关系陈述句是否表达同一关系的编码器f_θ：

$$p(l = 1|\mathbf{r}, \mathbf{r}^{'}) = \frac{1}{1 + e^{f_\theta(\mathbf{r})^\top f_\theta(\mathbf{r}')}} \tag{21}$$

$$\mathcal{L}(\mathcal{D}) = -\frac{1}{|\mathcal{D}|^2} \sum_{(\mathbf{r},e_1,e_2)\in\mathcal{D}} \sum_{(\mathbf{r}',e_1',e_2')\in\mathcal{D}} \alpha \mathbf{log} p(l = 1|\mathbf{r}, \mathbf{r}^{'}) + (1 - \alpha)\mathbf{log}(1 - p(l = 1|\mathbf{r}, \mathbf{r}^{'})) \tag{22}$$

其中，$l = 1$表示$\mathbf{r}$与$\mathbf{r}'$表示表示同一种关系，否则表示不同关系。$\alpha = \delta_{e_1,e_1'}\delta_{e_2,e_2'}$，$\delta_{e,e'}$为克罗内克函数，当且仅当$e = e'$时为1，否则为0。为了避免模型只是机械地记忆目标实体，而忽略了句子的语义，作者以概率β将目标实体随机替换为空白符[**BLANK**]。在如何标记句子目标实体问题上，作者采用了与基于GPT的原型网络相同的方法：(1)在目标实体两侧添加标记符；(2)位置嵌入。同时探索了如何从BERT的输出中得到固定长度的关系表示向量：(1)利用BERT原始的[**CLS**]；(2)拼接两个目标实体的池化表示；(3)在目标实体两侧添加标记符的基础上，拼接标记符[$\mathbf{E1}_{start}$]与[$\mathbf{E2}_{start}$]作为最终的关系表示向量。由于数据集过大，不可能比较所有的$\mathbf{r}$与$\mathbf{r}'$。作者采用了噪声对比估计训练方法(noise-contrastive estimation)，将所有包含同一对实体的关系陈述句视为正例对，从所有关系陈述句中随机选取一对句子或者选取只共享其中一个实体的句对构建负例对。最终，与BERT相似，作者平衡两种损失函数对模型进行无监督训练：

$$\mathcal{L} = \mathcal{L}_{match} + \lambda \mathcal{L}_{MLM} \tag{23}$$

- **词-类分布特征网络(Distributional Signatures(Bao et al., 2020))**通过学习在任务间具有一致性的词对类的分布特征来迁移任务间共享的元知识，同时根据词对类的重要程度构造句子表示，避免池化带来的信息丢失。模型分为两个部分，一是注意力权重生成器，另一个是用于分类的任务特定的岭回归器。权重生成器的目标是根据句子中词的分布特征生成词的重要程度。作者选用一元模型(Unigram)作为统计特征，增强对词替换扰动的鲁棒性。由于高频词通常不包含有用信息，为了降低高频词权重，增大低频词权重，作者度量了通用的词-词表重要程度：

$$s(x_i) = \frac{\varepsilon}{\varepsilon + P(x_i)} \tag{24}$$

其中$\varepsilon = 10^{-3}$，x_i是句子x的第i个词，$P(x_i)$是词x_i在整个元训练集上的一元模型似然。

同时，在支持集中相对具有辨识度的词，对于问题集可能也相对具有辨识度。因此，作者度量了特定的词-类别重要程度：

$$t(x_i) = H(P(y|x_i))^{-1} \tag{25}$$

其中，$H(\cdot)$表示熵操作，$P(y|x_i)$通过一个正则线性分类器在支持集上的估计得到。

考虑到这两种统计特征信息互补，且存在一定的噪声。作者通过BiLSTM将两者融合$h_i = \mathbf{BiLSTM}([s(x_i); t(x_i)])$，最终得到词$x_i$的权重($v$是可学习的元参数)：

$$\alpha_i = \frac{e^{v^\top h_i}}{\sum_j e^{v^\top h_j}} \tag{26}$$

在权重生成器的基础上，根据支持集构建岭回归器对问题集样本进行分类。作者首先根据词的权重，构建句子表示：

$$\phi(x) = \sum_i \alpha_i f_\theta(x_i) \tag{27}$$

然后，通过对支持集的拟合构建岭回归器(闭式解避免了梯度的二次迭代)：

$$\mathcal{L}_{RR}(\mathbf{W}) = \|\Phi_S \mathbf{W} - Y_S\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \tag{28}$$

$$\mathbf{W} = \Phi_S^\top (\Phi_S \Phi_S^\top + \lambda I)^{-1} Y_S \tag{29}$$

其中，$\Phi_S \in \mathbb{R}^{NK \times d}$表示整个支持集，$Y_S \in \mathbb{R}^{NK \times N}$表示独热标签，$I$为单位矩阵。

根据得到的岭回归器，对问题集样本进行分类：

$$\hat{Y}_Q = a\Phi_Q \mathbf{W} + b \tag{30}$$

其中，$a \in \mathbb{R}^+, b \in \mathbb{R}$为通过元学习得到的用于校正岭回归器参数的元参数。

最终，通过计算预测值与真实值之间的交叉熵损失训练整个模型。

4 稀缺资源小样本关系分类

当前元学习方法假设模型处理的任务服从同一分布。但在真实场景中，模型所遇到的新任务可能并不满足这一假设。其次，尽管在元测试阶段(Meta-Test)，元学习器只需要少量的监督数据，但在元训练阶段(Meta-Train)，训练元学习器所需要的监督数据依然很庞大，例如，FewRel数据集中每个类别700个样本。在一些领域，比如医疗、金融领域，获取元数据(Meta-Data)是十分困难的。直觉上，如果一些类别的样例很少，同领域的其它类别的样例也不足以构建一个足够大的数据集用以训练元学习器(Geng et al., 2020b)。因此，为了使元学习器能够在这些领域中发挥作用，研究者们从不同的角度提出了不同的解决方法。

4.1 小样本领域适应

Gao et al. (2019b)在FewRel数据集的基础上提出了FewRel2.0数据集。作者以包含大量生物医学文献的PubMed作为数据库，以UMLS作为知识库，利用FewRel1.0数据集的构建方法，构建了一个包含25种关系，每种关系100个样本的生物医学领域的数据集。FewRel2.0沿用了FewRel1.0的训练集，但是以新数据集为测试集，以此探究元学习模型从高资源领域向低资源领域适应的问题。同时，文章提出利用BERT序列分类模型解决此问题，在表现上远远超越了基于对抗的领域适应方法。

Wang et al. (2019a)在预训练语言模型的基础上，结合知识嵌入模型(KE)，将知识图谱中的事实知识融入预训练语言模型，提出了开普勒模型(KEPLER)。作者利用预训练语言模型RoBERTa，将句子中目标实体的文本表示与整个句子编码到统一的语义空间中，在预训练过程中联合优化知识嵌入模型与掩码语言模型。以KEPLER模型作为原型网络的编码器，整个网络在FewRel2.0数据集上取得了最优的表现。

4.2 小-小样本学习

Geng et al. (2020b)通过远程监督和人员筛选的方法，构建了一个新的中文医疗健康领域的小样本关系分类数据集TinyRel-CM，以探索在限制了元数据情况下的小样本学习(Few-few-shot Learning)。数据集包含27种4个实体间的二元关系，每种关系50个样本。作者根据实体类别将其分为6个部分，其中1个作为测试集，其余5个作为训练集，构建了6个任务。为了解决元训练数据不足的问题，作者提出了利用内部支持与跨领域支持的元学习框架MICK。该框架除了对问题集进行分类外，还对支持集进行了分类，以挖掘支持集内部的知识。此外，作者利用跨领域关系分类数据集对小样本任务进行数据增强。

Gao et al. (2020)提出滚雪球网络，一种新的自举方法，利用现有关系的语义知识来挖掘新关系的样本。作者利用关系孪生网络，基于现有关系分类数据集学习样例间的关系相似度度量。在此基础上，给定一个新关系及其少量标记样本，使用关系孪生网络从无标记语料库中累计可靠样本。然后利用这些样本训练关系分类器，提高分类器对新关系的新样本的泛化能力。

Method	FewRel1.0				FewRel2.0			
	5-1	5-5	10-1	10-5	5-1	5-5	10-1	10-5
Distributional Signatures	67.10	83.53	-	-	-	-	-	-
ProtoNet(CNN)	74.52	88.40	62.38	80.45	35.09	49.37	22.98	35.22
LM-ProtoNet	76.60	89.31	65.31	82.10	-	-	-	-
ProtoNet(GPT)	81.40	92.11	72.51	86.03	-	-	-	-
MLMAN	82.98	92.66	73.59	87.29	-	-	-	-
Matching The Blank	**93.86**	**97.06**	**89.20**	**94.27**	-	-	-	-
ProtoNet-ADV(CNN)	70.28	84.63	56.34	74.67	42.21	58.71	28.91	44.35
ProtoNet(BERT)	80.68	89.60	71.48	82.89	40.12	51.50	26.45	36.93
ProtoNet(RoBERTa)	85.78	95.78	77.65	92.26	64.65	82.76	50.80	71.84
ProtoNet(KEPLER)	88.30	95.94	81.10	92.67	66.41	**84.02**	51.85	**73.60**
BERT-PAIR	88.32	93.22	80.63	87.02	**67.41**	78.57	**54.89**	66.85
RoBERTa-PAIR	89.32	93.70	82.49	88.43	66.78	81.84	53.99	70.85
KEPLER-PAIR	90.31	94.28	85.48	90.51	67.23	82.09	54.32	71.01

Table 1: 小样本关系分类算法在常规和跨领域设定下的准确率。N-K表示N-way K-shot小样本设定。Distributional Signatures论文只发布了验证集上的结果。-ADV表示对抗训练。-PAIR表示序列分类模型。部分结果引用自Wang et al. (2019a)。

5 当前技术挑战与未来研究趋势

5.1 当前小样本关系分类的技术挑战

当前小样本关系分类的研究主要集中在同领域任务间的知识迁移，且依然需要庞大的元数据训练元学习器。但这个利用一个领域大量元数据训练出的元学习器很难直接应用到其它领域。尽管，研究者们利用大型预训练语言模型去解决这个问题(Gao et al., 2019b; Wang et al., 2019a)，但是并没有显式地用到目标领域的信息。因此，这些方法实际上是领域泛化的方法。

从领域适应的角度来看，我们将元学习视为以(D^{train}, h^*)为训练样本对的监督式机器学习，其处理的基本单位不再是样本x而是任务T。目前，小样本关系分类都是同构迁移学习，因此源领域与目标领域任务的特征空间相同$\mathcal{T}_S = \mathcal{T}_T$，任务的分布不同$p(T_S) \neq p(T_T)$。但无论是源领域还是目标领域，其最终目的都是学习一个对应于任务T的基学习器h^*，即两个领域的元任务(Meta-Task)相同。因此，小样本领域适应实际上应称为元学习领域适应，其本质是将元学习器从源领域适应到目标领域。但是，如果我们希望利用传统机器学习中的领域适应思想来解决元学习领域适应问题，需要面对两个挑战：

- 如何获取目标领域的任务?
  在传统领域适应中，为了将模型适应到目标领域，需要目标领域的样本(无论有无标签)。对应元领域适应，则需要目标领域的任务。由于元训练集与元测试集类别互斥，因此，目标领域的任务是未知的。如何从目标领域的无标注样本中构建合理的任务，是元学习领域适应的第一个挑战。一种最直观的方法是对目标领域无标签数据进行聚类，核心问题在

于特征的抽取。从Table 1发现，在源领域训练的元学习器，虽然在目标领域数据集上的表现有大幅下降，但也有一定的效果。因此，可以利用源领域的元学习器辅助目标领域聚类。Cong et al. (2020)从对抗训练的角度出发，通过最小熵原理保证目标领域的聚类效果。

- 如何抽取任务特征?
在传统领域适应中，源领域与目标领域的输出空间相同，但是输入的分布不同，一种有效的方法是抽取领域无关的样本特征。尽管，有研究者通过对抗训练，抽取样本层面的领域无关特征来解决元领域适应问题(Gao et al., 2019b)。但在元学习模型处理的基本单位为任务。一个任务并不只包含样本这一个属性。任务中类别之间的相似度，也决定了这个任务的难易程度。因此，如何合理地表达一个任务的特征是元学习领域适应的第二个挑战。同时，当前基于度量的元学习方法本质上是在抽取同领域任务间的通用特征，如果在此基础上同时抽取领域无关的特征，如何保证最终抽取的特征的辨识度能够满足分类需要也有待解决。在保证集合无序性条件下，一种简单的获取任务特征的方法是统计法，如对支持集向量逐元素取均值、求和、求积、求几何平均或取最值(Edwards and Storkey, 2017; Li et al., 2019; Oreshkin et al., 2018)。为了抓取任务中的类别特征及样本数量，Lee et al. (2020)则采用更高阶的统计特征，如方差、偏度和峰度，并对DeepSets(Zaheer et al., 2017)进行了改进。除此之外，根据支持集向量构造无向图，通过图嵌入方法也能获取任务特征。

5.2 未来的研究趋势

- 多模态多领域泛化
无论是从领域适应的角度，还是从小-小样本学习的角度，解决单个领域元数据不足的方法都是迁移其它领域的知识。领域适应方法从单领域对单领域的适应方向解决问题，但需要我们获取目标领域的任务。小-小样本学习从数据增强的角度，直接利用多个领域的小样本关系数据集。但从领域泛化的角度出发，训练一个可以从多领域泛化到多领域的元学习器，就避免了获取大量单个领域任务或样本的麻烦。尽管每个领域的元训练集样本量不大，但是多个领域合成的元训练集在一定程度上也满足了元学习器的训练要求(Guo et al., 2019; Triantafillou et al., 2019)。此外，除了迁移同构领域之间的知识，异构领域可能包含更多的监督信息。利用多模态信息训练元学习器也能缓解单个领域元训练集不足的问题。

- 预训练语言模型压缩
预训练语言模型被证明很适合处理小样本学习任务(Brown et al., 2020)。但是，庞大的参数量以及所需的算力，限制了其在一些线下场景的应用。而且，随着参数量的降低，其在小样本任务上效果也会出现下降。如何在不损失模型效果的情况下，压缩模型的大小，是未来的一个发展方向。

- 更合理的小样本学习设定
目前大部分小样本关系分类模型的本质是元学习在极端小样本设定下的应用(N-way K-shot)。一方面，从定义上来讲，小样本问题并不等同于元学习问题。另一方面，在真实场景中，任务的类别数N与其包含的样本数K并不是固定的(Lee et al., 2020)。近来，有研究者发现最朴素的微调方法，在小样本任务上超越了元学习方法(Chen et al., 2019; Chen et al., 2020)，也有研究者分别从理论与实验的角度证明了学习一个好的表示对小样本任务至关重要(Tian et al., 2020; Du et al., 2020)。因此，元学习方法并不是解决小样本问题的唯一出路。如何确立更接近真实场景的小样本学习设定也需要进一步研究。

6 总结

在这篇文章中，我们系统地梳理了小样本关系分类算法，从度量方法上，将现有方法分为基于原型的方法和基于分布式表示的方法。从是否利用额外信息的角度，将现有方法分为预训练式与非预训练式。基于原型的方法主要从特征抽取器的角度入手，根据小样本分类的特点对特征抽器作特定地设计。基于分布式的方法从句子层面和词的层面建模各自的分布表示。此外，本文介绍了稀缺资源场景下的小样本关系分类任务，指出当前用于这些任务的方法在一些应用场景的局限性。最后，针对这些局限性，展望了小样本关系分类未来的发展方向。

第十九届中国计算语言学大会论文集，第363页-第375页，海南，中国，2020年10月30日至11月1日。

致谢

感谢各位匿名评审给出的意见与建议。本论文受北京市自然科学基金项目(4192057)资助。

参考文献

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. 2016. Learning to learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, pages 3981–3989.

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

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *International Conference on Learning Representations*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Wei-Lun Chao, Han-Jia Ye, De-Chuan Zhan, Mark Campbell, and Kilian Q Weinberger. 2020. Revisiting meta-learning as supervised learning. *arXiv preprint arXiv:2002.00573*.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. 2019. A closer look at few-shot classification. In *International Conference on Learning Representations*.

Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. 2020. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*.

Xin Cong, Bowen Yu, Tingwen Liu, Shiyao Cui, Hengzhu Tang, and Bin Wang. 2020. Inductive unsupervised domain adaptation for few-shot classification via clustering. *arXiv preprint arXiv:2006.12816*.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. 2020. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*.

Markus Eberts. 2019. *Relation Extraction with Attention-based Transfer Learning*. Ph.D. thesis, Hochschule RheinMain, FB Design Informatik Medien, Informatik.

Harrison Edwards and Amos J. Storkey. 2017. Towards a neural statistician. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Miao Fan, Yeqi Bai, Mingming Sun, and Ping Li. 2019. Large margin prototypical network for few-shot relation classification with fine-grained features. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2353–2356, New York, NY, USA. Association for Computing Machinery.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019a. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6407–6414.

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

Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Neural snowball for few-shot relation learning. In *Proceedings of AAAI*.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3895–3904.

Ruiying Geng, Binhua Li, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020a. Dynamic memory induction networks for few-shot text classification. *arXiv preprint arXiv:2005.05727*.

Xiaoqing Geng, Xiwen Chen, and Kenny Q Zhu. 2020b. Mick: A meta-learning framework for few-shot relation classification with little training data. *arXiv preprint arXiv:2004.14164*.

Yunhui Guo, Noel CF Codella, Leonid Karlinsky, John R Smith, Tajana Rosing, and Rogerio Feris. 2019. A new benchmark for evaluation of cross-domain few-shot learning. *arXiv preprint arXiv:1912.07200*.

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. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. *arXiv preprint arXiv:2004.03186*.

Hae Beom Lee, Hayeon Lee, Donghyun Na, Saehoon Kim, Minseop Park, Eunho Yang, and Sung Ju Hwang. 2020. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. In *ICLR*.

Ke Li and Jitendra Malik. 2016. Learning to optimize. *arXiv preprint arXiv:1606.01885*.

Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. 2019. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10.

Li Fei-Fei, R. Fergus, and P. Perona. 2006. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, page 1003–1011, USA. Association for Computational Linguistics.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2017. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*.

Tsendsuren Munkhdalai and Hong Yu. 2017. Meta networks. *Proceedings of machine learning research*, 70:2554–2563.

Abiola Obamuyide and Andreas Vlachos. 2019. Model-agnostic meta-learning for relation classification with limited supervision. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5873–5879, Florence, Italy, July. Association for Computational Linguistics.

Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731.

Victor Garcia Satorras and Joan Bruna. 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4077–4087. Curran Associates, Inc.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. 2020. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. 2019. Meta-dataset: A dataset of datasets for learning to learn from few examples.

Joaquin Vanschoren. 2018. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*.

Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *NIPS*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019a. Kepler: A unified model for knowledge embedding and pre-trained language representation.

Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. 2019b. Generalizing from a few examples: A survey on few-shot learning. *arXiv preprint arXiv:1904.05046*.

Zihao Wang, Kwunping Lai, Piji Li, Lidong Bing, and Wai Lam. 2019c. Tackling long-tailed relations and uncommon entities in knowledge graph completion. 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 250–260, Hong Kong, China, November. Association for Computational Linguistics.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. *CoRR*, abs/1410.3916.

Yuxiang Xie, Hua Xu, Jiaoe Li, Congcong Yang, and Kai Gao. 2020. Heterogeneous graph neural networks for noisy few-shot relation classification. *Knowledge-Based Systems*, 194:105548.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. 2017. Deep sets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3391–3401. Curran Associates, Inc.