# 面向低资源场景的实体知识获取研究综述

徐道柱[1]，赵凯琳[2]，康栋[3]，马超[1]，冯禹铭[2]，李紫宣[2]，弋步荣[3]，靳小龙[2]
1. 西安测绘研究所，陕西 西安 710054；
2. 中国科学院计算技术研究所，北京 100086；
3. 航天恒星科技有限公司，北京 100089

**摘要**
实体获取是信息抽取的核心任务。近年来，在大数据训练模型的趋势下，深度学习在实体获取任务上取得了成功。但在自然环境等领域中，地形、灾害等类型的实体样本或者标注样本很少，而且对无标签样本进行标注又耗时费力。因此，面向低资源场景的实体获取逐渐受到关注，该任务被称作低资源实体获取或小样本实体获取。系统地梳理了当前低资源实体获取的相关工作，具体来说介绍了基于元学习、基于多任务学习和基于提示学习这3类方法的研究现状；总结了目前常用的低资源实体获取数据集和代表性模型在这些数据集上的实验结果；对低资源实体获取的方法进行了总结与分析；总结了低资源实体获取的挑战，并展望了未来发展方向。

**关键词**
实体获取；低资源场景；小样本学习

中图分类号：TP18 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023079

# *Survey on entity extraction for low-resource scenarios*

**XU Daozhu[1], ZHAO Kailin[2], KANG Dong[3], MA Chao[1], FENG Yuming[2], LI Zixuan[2], YI Burong[3], JIN Xiaolong[2]**
1. Xi'an Research Institute of Surveying and Mapping, Xi'an 710054, China
2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100086, China
3. Space Star Technology Co., Ltd., Beijing 100089, China

***Abstract***
Entity extraction is an essential task in information extraction. In recent years, under the trend of training model with big data, deep learning has achieved success in entity extraction. However, in the fields such as natural environment, there are very few entity samples or labeled samples of terrain, disasters and other types, and labeling those unlabeled samples is time-consuming and laborious. Therefore, entity extraction for low-resource scenarios has gradually attracted more and more attention, which is called low-resource entity extraction or few-shot entity extraction. This paper systematically combs the current approaches of low-resource entity extraction. It introduces the research status of three types of methods: meta-learning based, multi-task learning based, and prompt learning based. Next, the paper summarizes the low-resource entity

extraction datasets and the experimental results of the representative models on these datasets. In the following, the current low-resource entity extraction approaches are analysed. Finally, this paper summarizes the challenges of low-resource entity extraction and discusses the future research direction in this field.

***Key words***
entity extraction, low-resource scenarios, few-shot learning

## 0 引言

实体获取是指从非结构化文本中抽取特定类型的实体，例如在自然环境领域中抽取地形、灾害和地名等实体类型。在“2月28日在四川甘孜州泸定县发生4.8级地震，该县由于处于青藏高原与四川盆地交界处，经常发生地震”这句话中，“四川甘孜州泸定县”和“地震”分别为地名和灾害类型的实体，“青藏高原”和“四川盆地”为地形类型的实体。实体获取是信息抽取的核心任务，直接影响信息检索、知识问答、机器翻译、知识图谱构建等下游任务。近年来，随着大数据时代的到来、知识图谱以及自然语言处理等领域的快速发展，基于深度学习的实体获取技术已取得长足进步。但是，随着新的应用场景和应用领域的不断出现，数据的来源和渠道不断拓宽，缺少高质量标注数据成为构建深度实体获取模型的主要挑战之一。例如在自然环境领域中，灾害、地形类型的实体数据或者标注数据稀缺，而对无标签数据进行标注将会消耗大量的时间和人力。因此，如何在低资源场景下进行实体获取，逐渐受到人们的关注。该任务叫作低资源实体获取，也被称为小样本实体获取。

低资源实体获取旨在从含有少量样本的实体类型中学习到实体获取模型。目前，低资源实体获取一般依照N-way K-shot的形式进行构建，即一个N-way K-shot任务包含一对支持集和查询集。支持集包含$N$个实体类别，每个类别包含$K$个带标签样本，查询集包含与支持集相同的$N$个实体类别，每个类别包含$Q$个待预测样本。训练阶段和测试阶段均包含多个N-way K-shot任务，训练只在含有大量标注样本的源域上进行，测试只在含有少量标注样本的目标域上进行。**图1**展示了一个在自然环境领域中获取低资源实体的实例。

虽然实体获取经历了长时间的研究，但是低资源实体获取在近几年才被提出并逐渐成为热门问题。面对标注数据稀缺的难题，低资源实体获取问题的解决思路如下：在大量源域数据上得到一个预训练模型，通过迁移学习将某些知识迁移到目标域上，并进行快速泛化。根据不同的迁移学习方法，本文将目前的工作分为3类：基于元学习的低资源实体获取方法、基于多任务学习的低资源实体获取方法和基于提示学习的低资源实体获取方法。基于元学习的方法在多个低资源实体获取任务上学习元知识（模型的初始参数或者超参数等），使模型快速适应新的小样本任务。基于多任务学习的方法通过拆分主任务或

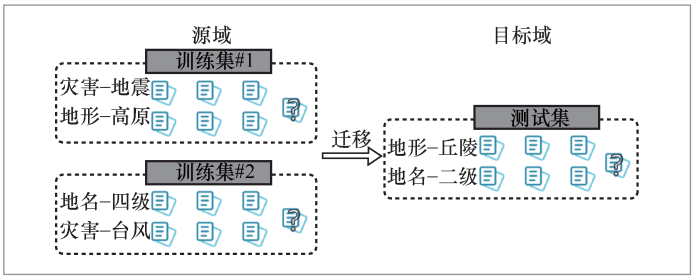


图1 低资源实体获取实例

者引入辅助任务来挖掘任务之间的联系，从而提升低资源实体获取的效果。基于提示学习的方法为预训练语言模型设计合适的提示，从而提高模型在小样本任务上的泛化能力、减缓过拟合问题。这3类模型均在低资源实体获取任务上取得了先进的成果。

本文首先从基于元学习、基于多任务学习和基于提示学习这3种方法介绍低资源实体抽取的工作进展。其次，总结了低资源实体获取的基准数据集以及代表性方法在这些数据集上的效果。然后，对这3种方法的研究现状以及优缺点进行了总结。最后，梳理了低资源实体获取目前面临的挑战，并展望了未来发展方向。

## 1 基于元学习的方法

小样本学习中的元学习方法的核心思想是通过双层次优化过程使元学习器在多个先验任务上学习到任务无关的元知识，最终实现在新任务上的泛化[1]。基于元学习的低资源实体获取的通用流程如下：首先，在源域上构建多个实体获取任务，每个任务包含相应的支持集和查询集；然后，使用元学习算法从源域的大量任务中学习到元知识，这些元知识可以提高模型在目标域上的性能；最后，在目标域上微调模型以适应目标域的数据和任务。该模型利用源域上学习到的知识来快速适应目标域的任务，从而获得更好的性能。基于元学习的低资源实体获取方法可以分为4类：基于度量的元学习方法、基于优化的元学习方法、基于记忆的元学习方法和基于信息增强的元学习方法。

基于度量学习的元学习方法通过距离函数学习到泛化性更强的编码器，从而提升低资源实体获取的效果。2019年，Fritzler等[2]提出了一种基于原型网络[3]的方法，即Proto模型，该方法将同一类型实体表示的平均值作为该类型的原型，然后基于实体类型原型表示和词语表示的相似度进行实体分类。Yang等[4]将在源域上训练的监督实体获取模型作为特征提取器，提出了一个基于最近邻学习和结构化推理的实体获取模型，分别是NNShot模型、StructShot模型。针对低资源实体获取任务的标签依赖问题，Hou等[5]提出了L-TapNet+CDT模型，如**图2**所示。该模型设计了一种分数转移的迁移机制，使得条件随机场（conditional random field，CRF）[6]能够适用源域和目标域的实体类型集合不一致的场景。此外，该模型将标签语义的表示融入类型原型表示中，增强了低资源场景下实体的原型表征，从而提升低资源实体获取的效果。但这些方法只从源域学习特定类型的语义特征和中间表示，这会影响目标域的泛化性，导致性能次优。为此，Das等[7]提出了一种新的对比学习技术，即CONTaiNER模型，用于优化低资源实体获取的序列间分布距离，从而提升了模型在目标域上的泛化性。上述方法通过标记分类来解决低资源实体获取问题，忽略了实体边界信息，不可避免地会受到大量非实体标记的影响。针对这个问题，Wang等[8]提出了一种开创性的基于跨度的原型网络SpanProto，该网络通过两阶段方法处理低资源实体获取，包括跨度提取和分类，如**图3**所示。

基于优化的元学习方法，通过设计更好的优化策略或优化器来增强元学习器捕捉不同任务的共性能力，最终达到更好的迁移效果。Li等[9]设计了一种显式区分任务相关参数和任务无关参数的元学习模型，在该模型的优化过程中，内部循环更新任务相关参数，外部循环通过二阶导数更新任务无关参数。为了更好地实现领域

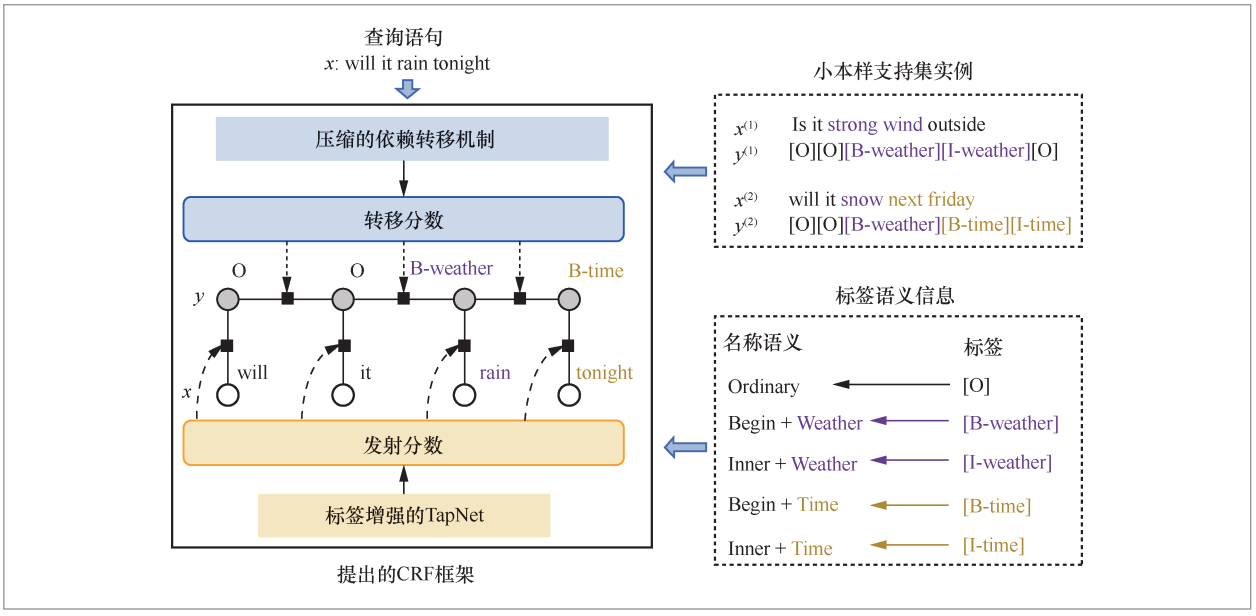


图 2 L-TapNet+CDT 模型结构 [5]

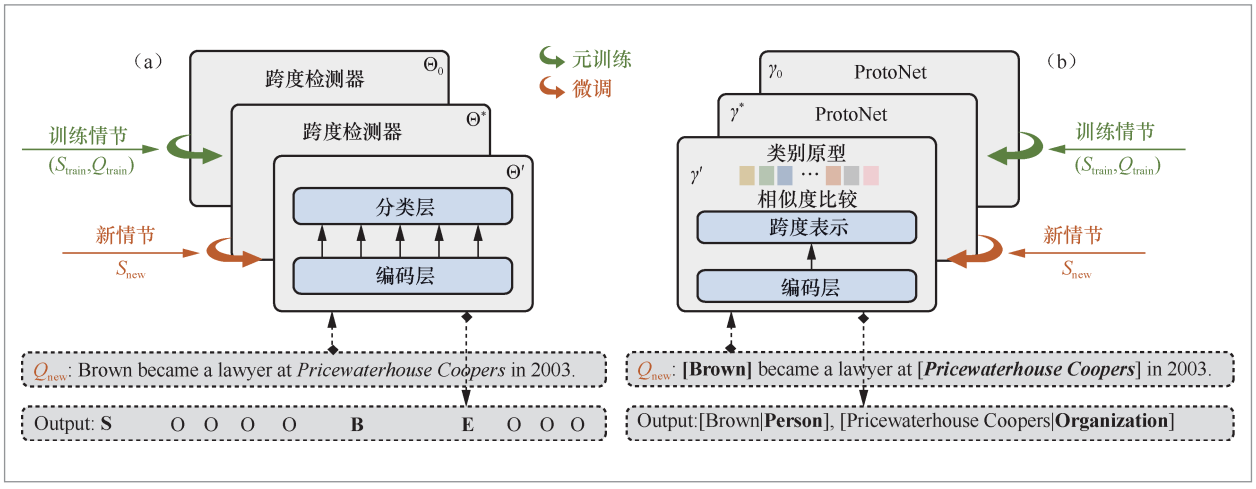


图 3 SpanProt 模型架构 [8]

相关和领域无关特征的解耦，Li等[10]在低资源实体获取任务上第一次引入了模型无关的元学习方法（model agnostic meta learning，MAML）[11]和领域对抗训练方法，使模型能够在多个源域的混合数据上训练。具体地，该模型设置了两个优化目标，使编码器提取的特征在丰富性和领域无关性上达到平衡。

基于记忆模块的元学习方法通过设计外部记忆模块存储源域样本的表示，并利用注意力机制引入源域信息，从而增强目标域样本的特征表示。为了将源域中实体类型知识更好地迁移到目标域的新类型上，Zhang等[12]设计了一种能够适用于零样本场景的记忆增强模型MZET，该模型利用外部记忆模块存储源域中的旧实体类

型的表示，并通过对旧类型表示和新类型表示之间的相关性进行建模，将词语和新实体类型的相似度度量转换成词语和旧实体类型的相似度度量。

基于信息增强的元学习方法通过引入外部信息来增强原型表征，从而提升基于原型网络的低资源实体获取效果。Ji等[13]提出了一种基于实体级原型网络的方法，从而解决实体标签依赖和语义空间中原型过于接近的问题。Wen等[14]引入了句子级别的语义信息来提高模型的鲁棒性。然而，大多数工作是基于标记的相似性为每个标记分配一个标签，忽略了实体的完整性。Wang等[15]提出的ESD模型是一种增强的基于跨度的分解方法，将低资源实体获取任务转化为测试查询和支持实例之间的跨级匹配问题。而Ma等[16]提出了一种分解元学习方法DecomMetaNER，该方法通过使用元学习连续处理小样本跨度检测和小样本实体类型分类来解决低资源实体获取问题，将小样本跨度检测视为序列标记问题，并通过引入MAML算法来训练跨度检测器，以找到能够快速适应新的实体类型的模型初始化参数。Huang等[17]采用了类型描述增强的策略，提出了DFS-NER模型，该模型在实体类型原型表示上通过单词-单词级和单词-类型级的对比学习和胶囊网络进行信息增强，并且引入了一个由类型描述引导的掩码语言模型学习目标，从而更好地利用实体类型的语义信息。

总的来说，由于在小样本任务上建立了鲁棒高效的元学习器，基于元学习的方法缓解了在目标域上过拟合的问题，但效果受限于目标域样本质量。元学习方法在低资源实体获取任务中具有广泛的应用前景。随着研究的不断深入，该领域的发展趋势将更加注重模型的普适性和可解释性，引入跨模态元学习和多任务元学习等新的研究技术，从而提升模型在低资源实体获取任务中的性能。

## 2 基于多任务学习的方法

由于目标域中实体类型的标注样本非常有限，使用单任务学习方式进行低资源实体获取的效果会受到很大影响。相比之下，任务之间的相关性可以帮助模型利用有限的数据更好地学习泛化特征，因此，多任务学习方式能够更好地挖掘任务之间的联系，从而提高每个任务的表现。此外，多任务学习还可以提高模型的鲁棒性，更好地适应不同领域和任务。

基于多任务学习的方法通常分为以下几个步骤：首先，设计合适的任务分解方式，将低资源实体获取任务分解成多个子任务或引入辅助任务；其次，设计共享编码器和与任务相关的私有模块，其中共享编码器用于提取通用特征，私有模块用于与学习任务相关的特征；然后，在域数据上进行模型的预训练，提高模型的泛化性能；最后，在目标域上进行模型微调，以适应目标域的数据和任务。

将低资源实体获取任务拆分成多个子任务，能够减少由于源域和目标域实体类型集合不一致而产生的迁移偏差。拆分后的子任务通常包括针对多个实体类型的词语级别匹配任务、实体边界检测任务和实体分类任务。Bapna等[18]第一次将低资源实体获取任务建模成词语级别匹配任务，先通过LSTM编码句子表示，再拼接实体类型描述的表示，最后输入LSTM进行序列标注。在此基础上，Lee等[19]增加了注意力机制以更好地将实体类型信息融入词语表示，并使用CRF作为序列标注器。为了有效利用目标域的少量标注样本，Shah等[20]在对输入样本的词语进行编码时增加了注意力机制，以平衡标注样

本中不同词语的权重。Liu等[21]第一次提出了不需要额外资源的零样本跨领域适应实体获取模型MTL+MoEE，如**图4**所示。该模型联合实体边界检测任务和实体分类任务，同时引入了混合专家模型以平衡不同类型的实体表示，提升了模型的鲁棒性。

为了更好地将源域的知识迁移到目标域上，可以设计辅助任务来帮助低资源实体获取。针对零样本场景，Zhang等人[22]提出了一个多域数据混合、多阶段、多任务的训练框架，从任务、语言和领域3个维度划分知识，并探究了如何选择训练数据和微调任务问题。该框架引入了掩码语言模型任务和基于机器阅读理解的实体边界检测任务，增强了底层编码器的跨域表示能力，从而提升性能。

总的来说，基于多任务学习的方法可以通过多个学习目标的约束，在更充分的监督信号上学习到泛化表示能力更好的模型。该方法的优点在于，通过在多个相关任务之间进行约束，提供额外的训练信息，从而提高每个任务的表现。另外，设计一些与实体获取相关的辅助任务，可以帮助模型更好地将从源域获得的知识迁移到目标域上。然而，基于多任务学习的方法也存在一些问题。首先，辅助任务的增益效果难以提前估计，需要通过实验来确定哪些任务对主任务性能的提升效果较好；其次，在多任务学习过程中，需要平衡不同任务对主任务学习的影响，以免影响主任务的学习；最后，在多任务学习中，不同任务之间可能存在冲突，这需要对任务、模型架构等进行多次设计。

针对基于多任务学习的方法，未来的发展趋势包括更有效的任务设计、更好的数据利用、更好的领域适应能力以及更好的跨语言学习能力。如何设计更加有效的辅助任务，如何利用源域和目标域的信息来提高模型的泛化能力，如何在不同的领域之间进行迁移学习，如何将在一种语言上训练的模型迁移到另一种语言上，这些都将成为基于多任务学习的低资源实体获取方法的重要研究方向。

## 3 基于提示学习的方法

随着大规模预训练语言模型的发展，

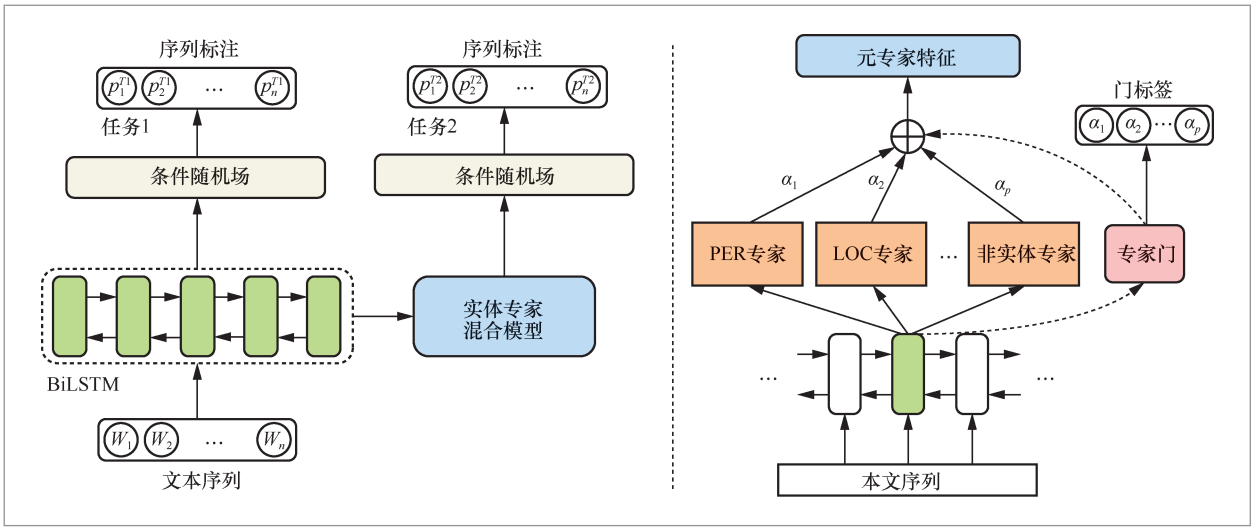


图 4 MTL+MoEE 模型结构 [21]

提示学习成为自然语言处理领域的新研究范式[23-24]。其核心思想是将下游任务转换成与预训练阶段形式相似的任务，从而最大化利用预训练语言模型的泛化能力。基于提示学习的方法的通用流程如下：首先，确定下游任务类型和目标领域，将下游任务改造为与预训练阶段形式相近的任务；其次，通过对目标域标注样本进行优化或选择提示，对预训练模型进行微调来进行目标域上的低资源实体获取。

近年来，低资源实体获取领域开始探索将该任务建模成Seq2Seq任务，并利用具有强大泛化能力的预训练模型来解决小样本问题。先前的方法依赖于源域和目标域之间相似文本特征进行知识迁移，而Cui等[25]提出了一种基于双向自回归变换器（bidirectional autoregressive transformers，BART）的提示学习模型Template-based BART，如图5所示。在训练阶段，该模型接收待预测样本作为输入，并为正确预测结果对应的提示输出产生最高的概率分数。在测试阶段，该方法需要枚举所有句子片段和所有实体类型提示的组合，以得到对某一句的预测结果。

预训练模型的自编码任务通常要填充句子中缺失的部分，在低资源实体获取任务中具有很大的潜力，可以更好地表示和抽取文本信息。然而，传统基于提示学习的方法存在搜索空间过大、提示设计困难等问题。针对这些问题，Ma等[26]提出了EntLM模型，该模型提出了一种新的适用于掩码语言模型的提示构造方式，这种方式使用原句子作为输入提示，然后选择某一实体类型的高频词语作为标签词，在标签词替换后得到输出提示，从而获得更好的文本表示。Huang等人[27]针对低资源实体获取任务中实体标签不足的问题，提出了COPNER方法，通过引入面向特定类别的词汇来进行对比学习和度量推理。

近来的一些研究表明，将实体获取任务建模成一些其他成熟领域的任务，能够提升低资源实体获取的效果。Sun等[28]发现将继承预测任务应用在小样本学习问题中，可以有效发挥预训练模型的能力，具体流程为将原句子和预测结果对应的提示一起输入预训练模型，无监督地选择存在继承关系、预测概率最高的一对作为下游任务的预测结果。Li等人[29]将低资源实体获取任务转化为文本继承预测任务，提出了基于提示学习的文本继承方法PTE，利用预训练语言模型对候选实体进行打分，从而得到最终的实体标签。Liu等人[30]则提出了QaNER方法，将实体获取问题转化为问答问题，并在已有的问答模型上添加提示学习，取得了更好的效果。

总的来说，基于提示学习的方法对低资源实体获取任务形式进行了改造，使其与预训练阶段保持一致，从而降低了微调

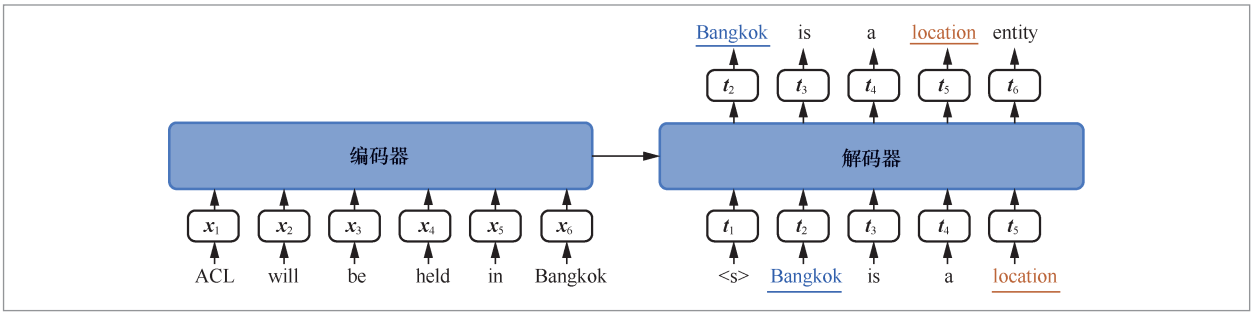


图 5 Template-based BART 模型结构 [25]

参数对目标域标注数据的需求，更充分地利用了预训练语言模型的泛化能力。然而，传统的提示学习方法需要枚举所有句子片段和所有实体类型提示的组合，导致搜索空间过大。最新的连续型提示学习的缺点在于提示的设计会影响学习效果，在小样本场景下缺少充分的监督信号去优化或筛选高质量的提示。

基于提示学习的低资源实体获取方法的未来发展方向是将提示学习与其他技术结合以提高实体获取的效果。例如，使用强化学习来设计有效的提示，或者使用元学习来针对不同的数据集和任务进行模型调整。此外，还可以使用其他更有效的方法，如图神经网络方法和注意力机制方法等，更好地表示和抽取文本信息。同时，可以提出更多新的提示学习范式，将实体获取问题转化成其他成熟领域的问题。

## 4 数据集与实验

在低资源实体获取中，常用的基准数据集有SNIPS、Few-NERD和CrossNER。此外，本文针对自然环境领域构建了一个低资源实体抽取数据集FewBE。接下来，对这4个数据集进行介绍，并对低资源实体获取方法在这些数据集上的实验结果进行总结和分析。

### 4.1 数据集

SNIPS数据集[31]是一个槽填充数据集，常用于实体获取任务。该数据集标注了39种细粒度实体类型，官方划分了7个不同的数据域，即GetWeather、PlayMusic、AddToPlaylist、BookRestaurant、SearchScreeningEvent、RateBook、SearchCreativeWork，分别简称为We、Mu、P1、Bo、Se、Re和Cr。该数据集通过留一法来构建低资源实体获取评测任务，具体而言，选择一个数据域用于测试，一个数据域用于验证，其他数据域用于训练，在不同数据域上总共划分出7个任务。

CrossNER数据集[32]是一个完全标注的跨域实体获取数据集，涵盖5个不同的领域，分别为政治、自然科学、音乐、文学和人工智能，并为不同的领域提供专门的实体类别。此外，该数据集还提供了与各个领域相关的4个语料库，包括OntoNotes 5.0、CoNLL-03、GUM和WNUT-17。

Few-NERD数据集[33]是一个专门为低资源实体获取任务设计的数据集，也是当今最大的开源实体获取数据集之一，标注了8个粗粒度实体类型和66个细粒度实体类型。同时，官方还给出了两种不同的评测任务以及相应的数据划分，分别为Few-NERD-INTRA和Few-NERD-INTER。在Few-NERD-INTRA任务中，在源域上的训练集和在目标域上的验证集和测试集的实体类型属于不同粗粒度类型。在Few-NERD-INTER任务中，它们在粗粒度类型上有交集，在细粒度类型上无交集。由于采样的数据的类型数和样本数存在差异，这两个任务都有4种不同的设定，即5-way 1~2-shot、5-way 5~10-shot、10-way 1~2-shot和10-way 5~10-shot。

数据集FewBE是本文针对自然环境领域构建的低资源实体获取构建的数据集，包含地形、灾害和地名3种实体类型，其中地形包含高原、平原、盆地、山地和丘陵5类实体类型，灾害包含了台风、海啸、地震、暴雨洪涝、干旱和泥石流6类实体类型，地名包含了一级、二级、三级和四级行政区划4类实体类型，共计18类实体类型，其中每一类含有20个标注样本。与

Few-NERD相同，本文给出了两种评测任务和数据集划分方式，即FewBE-INTRA和FewBE-INTER。在FewBE-INTRA任务中，源域上的训练集包括灾害和地名两个大类，目标域上的测试集包括地形大类，源域与目标域在大类上没有交集。在FewBE-INTER任务中，它们在粗粒度类型上有交集，在细粒度类型上无交集。由于采样的数据的类型数量和样本数量较少，这两个任务都有2种不同的设定，即3-way 1-shot和5-way 1-shot。

### 4.2 实验结果

针对SNIPS数据集，本文选取了L-TapNet+CDT、ESD和DFS-NER作为代表性模型。**表1**给出了各方法在SNIPS数据集上5-way任务设置下的性能对比。从**表1**可知，L-TapNet+CDT的平均F1值就可以达到70%左右，说明在该数据集上的低资源实体获取任务相比于其他两个数据集更简单。此外，随着低资源实体获取技术的不断发展，实验效果也在不断提升，但提升效果相比其他两个数据集较慢。近几年在这个数据集上的工作较少，遇到了瓶颈。从**表1**中同一模型的横向对比可以看出，模型在不同数据域上效果差异较大，例如DFS-NER在Bo和Mu领域相差24%，说明该模型不能同时适用于所有领域，泛化性有待提升。

针对CrossNER数据集，本文选取了L-TapNet+CDT、DecomMetaNER和SpanProto这3个代表性模型，其中SpanProto是目前的最先进模型。**表2**给出了这些模型在CrossNER数据集上5-way任务设置下的性能对比。通过不同模型间的纵向对比可以看出，近年来随着低资源实体获取技术的更新，新出的模型效果提升较大，但实验结果仍有很大的提升空间。从同一模型的横向对比可以看出，不同领域之间的实验结果差异较大，说明同一个模型对不同领域的普适性和泛化性较差，该结论与SNIPS数据集相似。

针对Few-NERD数据集，本文选取

表1 SNIPS 数据集低资源实体获取结果

| 设置 | 模型 | We | Mu | P1 | Bo | Se | Re | Cr | 平均F1值 |
|---|---|---|---|---|---|---|---|---|---|
| 1-shot | L-TapNet+CDT[5] | 71.53% | 60.56% | 66.27% | 84.54% | 76.27% | 70.79% | 62.89% | 70.41% |
| | ESD[15] | 78.25% | 54.74% | 71.15% | 71.45% | 67.85% | 71.52% | 78.14% | 70.44% |
| | DFS-NER[17] | 77.61% | 65.21% | 72.39% | 89.31% | 78.11% | 72.65% | 67.50% | 74.68% |
| 5-shot | L-TapNet+CDT[5] | 71.64% | 67.16% | 75.88% | 84.38% | 82.58% | 70.05% | 73.41% | 75.01% |
| | ESD[15] | 84.50% | 66.61% | 79.69% | 82.57% | 82.22% | 80.44% | 81.13% | 79.59% |
| | DFS-NER[17] | 80.42% | 76.81% | 84.52% | 90.02% | 86.79% | 78.32% | 84.81% | 83.10% |

表2 CrossNER 数据集低资源实体获取结果

| 设置 | 模型 | CoNLL-03 | GUM | WNUT-17 | OntoNotes 5.0 | 平均F1值 |
|---|---|---|---|---|---|---|
| 1-shot | L-TapNet+CDT[5] | 44.30% | 12.04% | 20.80% | 15.17% | 23.08% |
| | DecomMetaNER[16] | 46.09% | 17.54% | 25.14% | 34.13% | 30.73% |
| | SpanProto[8] | 47.70% | 19.92% | 28.31% | 36.41% | 33.09% |
| 5-shot | L-TapNet+CDT[5] | 45.35% | 11.65% | 23.30% | 20.95% | 25.31% |
| | DecomMetaNER[16] | 58.18% | 31.36% | 31.02% | 45.55% | 41.53% |
| | SpanProto[8] | 61.88% | 35.12% | 33.94% | 48.21% | 44.79% |

了在该数据集的开源网址上评测过的几个模型，如Proto、NNShot、StructShot、CONTaiNER、ESD和DecomMetaNER等，以及当前最先进的SpanProto模型，展示了这些模型在Few-NERD数据集上设置为1~2-shot的性能对比，见**表3**。根据实验结果可知，SpanProto相比于最早的低资源实体获取模型Proto，在每个任务上的提升均有30%~35%，说明近几年低资源实体获取技术取得了长足进步。此外，SpanProto模型在该数据集上的平均实验结果相比于SNIPS数据集来说较差，说明该数据集上的低资源实体获取任务更难，实验结果仍有较大的提升空间。

针对FewBE数据集，本文选取了一些小样本实体获取的开源模型进行效果评测，如Proto、NNShot、StructShot、DecomMetaNER和ESD等，**表4**给出了这些模型在FewBE-INTRA和FewBE-INTER两种测评任务上设置为3-way 1-shot和5-way 1-shot的性能对比。通过纵向对比可以看出，低资源实体获取方法在近年来取得了长足进步，该结论与Few-NERD数据集上的结论类似。通过横向对比可以看出，同一模型在INTRA和INTER两种数据集划分方式上的性能差异较大，表明了灾害、地名和地形实体类型之间存在较大差别。

## 5 挑战与展望

由于真实领域中某些类型的实体样本量或标注样本很少，而对样本进行标注会耗费大量人力和时间，近年来低资源实体获取逐渐获得关注。根据不同的迁移学习

表 3 Few-NERD 数据集低资源实体获取结果

| 模型 | INTRA(5-way) | INTRA(10-way) | INTER(5-way) | INTER(10-way) |
|---|---|---|---|---|
| Proto[2] | 20.76% | 15.05% | 38.83% | 32.45% |
| NNShot[4] | 25.78% | 18.27% | 47.24% | 38.87% |
| StructShot[4] | 30.21% | 21.03% | 51.88% | 43.34% |
| L-TapNet+CDT[5] | 25.81% | 18.02% | 41.44% | 36.80% |
| DFS-NER[17] | 35.41% | 20.31% | 48.03% | 34.38% |
| CONTaiNER[7] | 40.43% | 33.84% | 55.95% | 48.35% |
| ESD[15] | 36.08% | 30.00% | 59.29% | 52.16% |
| DecomMetaNER[16] | 49.48% | 42.84% | 64.75% | 58.65% |
| SpanProto[8] | 54.49% | 45.39% | 73.36% | 66.26% |

表 4 FewBE 数据集低资源实体获取结果

| 模型 | INTRA(3-way) | INTRA(5-way) | INTER(3-way) | INTER(5-way) |
|---|---|---|---|---|
| Proto[2] | 20.50% | 9.91% | 50.55% | 48.23% |
| NNshot[4] | 17.99% | 7.26% | 67.97% | 65.44% |
| StructShot[4] | 19.46% | 16.25% | 67.94% | 63.90% |
| DecomMetaNER[16] | 21.31% | 17.90% | 75.32% | 66.67% |
| ESD[15] | 25.48% | 20.65% | 74.21% | 69.04% |

方法，本文将低资源实体获取分为基于元学习、基于多任务学习和基于提示学习的方法。本文对以上3种方法进行了总结，比较了它们的优点和缺点，具体见**表5**。

总体来说，低资源实体获取方法已经取得了不错的进展。基于元学习的方法在大量任务上学习到能够提高模型泛化能力的元知识，因此不易在目标域过拟合，但是这类方法的性能不稳定，且严重依赖目标域的标注样本。基于多任务学习的方法由于多个优化目标的约束，泛化表征能力较强，但是严重依赖于辅助任务的质量，还需要平衡不同的辅助任务对主任务的影响程度。基于提示学习的方法利用预先设计好的提示，能够降低对标注数据的依赖，但是该类方法在小样本场景下的监督信号十分有限，并且严重依赖于提示的设计与选择。

## 5.1 挑战

通过对当前方法的优缺点分析，本文对低资源实体获取目前面临的挑战进行了总结。

（1）训练成本较高

目前的低资源实体获取方法，均需要在大规模源域数据集的大量任务上对模型进行预训练，而这需要花费大量的时间和计算资源。

（2）方法与任务结合不紧密

很多小样本方法聚焦在模型层面，没有与实体获取任务进行有效结合。在实体获取任务中，现有的方法往往忽略了不同类型实体之间的关联、实体与文本上下文之间的联系、先验的知识图谱等知识，而这些知识可能会对该任务的性能有所提升。

（3）对新的实体类型的适应性不足

在新领域构建实体获取模型，除了缺乏领域内的标注数据，可能会不断出现人为定义的新实体类型。然而，目前很多工作中的目标域是单一特定的，目标任务的实体类型集合是已知的，这些工作通过使用标签特定的解码框架来提升效果，但是无法泛化到新的实体类型和目标域上。

目前基于提示学习的方法由于设计了与任务相关的提示，在下游的低资源实体获取任务中无须引入额外的参数，在一定程度上缓解了训练成本高的问题。基于多任务学习的方法，需要针对低资源实体获取任务进行拆分或者引入相关的辅助任务，因此该类方法与实体获取任务结合相对紧密，但仍然忽视了隐含在实体获取中的知识，比如不同实体类型之间的关联。基于元学习的方法在大量任务上学习到了帮助实体获取的元知识，在新领域上泛化性较好，能够提高对新领域实体类型的适应性，但是在不同领域上的效果相差较大，仍然不能适用于所有的新领域。在当前的工作中，已有元学习和多任务学习、元学习和提示学习这两种结合方式，但是没有尝试多任务学习和提示学习、3种方法的结合方式。

**表 5　3 类低资源实体获取方法的优缺点**

| 方法 | 优点 | 缺点 |
| --- | --- | --- |
| 基于元学习的方法 | 不容易在目标域上过拟合 | 性能不稳定；<br>严重依赖目标域标注样本 |
| 基于多任务学习的方法 | 模型泛化表征能力强 | 依赖于辅助任务的选择与质量；<br>难以平衡不同辅助任务对主任务的影响程度 |
| 基于提示学习的方法 | 对标注数据依赖低 | 监督信号有限；<br>严重依赖于提示的选择 |

### 5.2 展望

通过对当前研究进展的梳理，本文对未来低资源实体获取的发展方向的展望如下。

（1）在数据和预训练模型层面，可以更好地利用知识图谱等先验知识，从而减少模型对大规模预训练数据、训练时间和计算资源的需求，降低模型的训练成本。随着ChatGPT[34]等大规模预训练模型的兴起，可以考虑利用这些模型卓越的小样本学习能力来提高在自然环境等低资源领域的实体获取效果。但在真实领域中，需要考虑应用场景、模型训练部署成本等因素，让这些大规模预训练模型真正落地。

（2）在任务层面，将小样本学习方法和实体获取任务结合得更紧密，在模型中更好地融入实体以及上下文的知识，比如不同类型实体之间的关联、实体与文本上下文之间的联系、实体文本中不同词之间的依赖关系等。

（3）在方法层面，针对不同方法有不同的展望。针对基于元学习的方法，设计更好的元学习器，以提高模型的普适性和可解释性。如何设计元学习器使其学习到更多、更有效的元知识，从而增强实体获取模型在不同新领域新实体类型上的泛化性，将是一个重要的研究方向。针对基于多任务学习的方法，设计更好的任务拆分方案或者设计更有效的辅助任务，还需要平衡不同任务对主任务学习的影响程度。针对基于提示学习的方法，设计更简单、有效的提示。在小样本场景下，需要利用有限的监督信号去优化或筛选高质量的提示。此外，可以尝试将这3种方法进行结合，以应对低资源实体获取中不同方面的挑战。

## 6 结束语

由于自然环境等真实领域中实体样本不足，低资源实体获取越来越受到人们的重视，并且在真实领域展现出良好的应用前景。近来，随着ChatGPT等一系列在低资源场景下具有高学习能力的大规模预训练模型的广泛应用，低资源实体获取也面临了一些机遇和挑战。在真实应用领域中，需要设计让大规模预训练模型真正落地的方案，在亟须解决的问题和场景中发挥作用。

## 参考文献:

[1] HOSPEDALES T, ANTONIOU A, MICAELLI P, et al. Meta-learning in neural networks: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5149-5169.

[2] FRITZLER A, LOGACHEVA V, KRETOV M. Few-shot classification in named entity recognition task[C]//Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. New York: ACM, 2019: 993-1000.

[3] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 4080-4090.

[4] YANG Y, KATIYAR A. Simple and effective few-shot named entity recognition with structured nearest neighbor learning[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 6365-

6375.

[5] HOU Y T, CHE W X, LAI Y K, et al. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 1381-1393.

[6] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. New York: ACM, 2001: 282-289.

[7] DAS S S S, KATIYAR A, PASSONNEAU R, et al. CONTaiNER: few-shot named entity recognition via contrastive learning[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 6338-6353.

[8] WANG J N, WANG C Y, TAN C Q, et al. SpanProto: a two-stage span-based prototypical network for few-shot named entity recognition[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 3466-3476.

[9] LI J, CHIU B, FENG S S, et al. Few-shot named entity recognition via meta-learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(9): 4245-4256.

[10] LI J, SHANG S, SHAO L. MetaNER: named entity recognition with meta-learning[C]//Proceedings of The Web Conference 2020. New York: ACM, 2020: 429-440.

[11] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning. New York: ACM, 2017: 1126-1135.

[12] ZHANG T, XIA C Y, LU C T, et al. MZET: memory augmented zero-shot fine-grained named entity typing[C]// Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg: International Committee on Computational Linguistics, 2020: 77-87.

[13] JI B, LI S, GAN S, et al. Few-shot named entity recognition with entity-level prototypical network enhanced by dispersedly distributed prototypes[EB]. arXiv preprint, 2022, arXiv: 2208.08023.

[14] WEN W, LIU Y B, LIN Q, et al. Few-shot named entity recognition with joint token and sentence awareness[J]. Data Intelligence, 2023, 5(3): 767-785.

[15] WANG P Y, XU R X, LIU T Y, et al. An enhanced span-based decomposition method for few-shot sequence labeling[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2022: 5012-5024.

[16] MA T T, JIANG H Q, WU Q H, et al. Decomposed meta-learning for few-shot named entity recognition[C]//Proceedings of Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg: Association for Computational Linguistics, 2022: 1584-1596.

[17] HUANG H N, FENG Y M, JIN X L, et al. DFS-NER: description enhanced few-shot NER via prompt learning and meta-learning[C]//Proceedings of 2022 IEEE/WIC/ACM International Joint Conference

on Web Intelligence and Intelligent Agent Technology (WI-IAT). Piscataway: IEEE Press, 2023: 796-803.

[18] BAPNA A, TÜR G, HAKKANI-TÜR D, et al. Towards zero-shot frame semantic parsing for domain scaling[C]// Proceedings of Interspeech 2017. ISCA: ISCA, 2017: 2476-2480.

[19] LEE S, JHA R. Zero-shot adaptive transfer for conversational language understanding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 6642-6649.

[20] SHAH D, GUPTA R, FAYAZI A, et al. Robust zero-shot cross-domain slot filling with example values[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 5484-5490.

[21] LIU Z H, WINATA G I, FUNG P. Zero-resource cross-domain named entity recognition[C]//Proceedings of the 5th Workshop on Representation Learning for NLP. Stroudsburg: Association for Computational Linguistics, 2020: 1-6.

[22] ZHANG Y, YANG Q. A survey on multi-task learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(12): 5586-5609.

[23] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners[EB]. arXiv preprint, 2021, arXiv: 2109.01652.

[24] ZHANG N, LI L, CHEN X, et al. Differentiable prompt makes pre-trained language models better few-shot learners[EB]. arXiv preprint, 2021, arXiv: 2108.13161.

[25] CUI L Y, WU Y, LIU J, et al. Template-based named entity recognition using BART[C]//Proceedings of Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 1835-1845.

[26] MA R T, ZHOU X, GUI T, et al. Template-free prompt tuning for few-shot NER[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2022: 5721-5732.

[27] HUANG Y, HE K, WANG Y, et al. Copner: contrastive learning with prompt guiding for few-shot named entity recognition[C]//Proceedings of the 29th International conference on computational linguistics. Gyeongju: International Committee on Computational Linguistics, 2022: 2515-2527.

[28] SUN Y, ZHENG Y, HAO C, et al. NSP-BERT: a prompt-based zero-shot learner through an original pre-training task: next sentence prediction[EB]. arXiv preprint, 2021, arXiv: 2109.03564.

[29] LI D, HU B, CHEN Q. Prompt-based text entailment for low-resource named entity recognition[EB]. arXiv preprint, 2022, arXiv: 2211.03039.

[30] LIU A T, XIAO W, ZHU H, et al. QaNER: prompting question answering models for few-shot named entity recognition[EB]. arXiv preprint, 2022, arXiv: 2203.01543.

[31] COUCKE A, SAADE A, BALL A, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces[EB]. arXiv preprint, 2018, arXiv: 1805.10190.

[32] LIU Z H, XU Y, YU T Z, et al. CrossNER: evaluating cross-domain named entity recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13452-13460.

[33] DING N, XU G W, CHEN Y L, et al. Few-NERD: a few-shot named entity recognition dataset[C]//Proceedings of the 59th Annual Meeting of

the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 3198–3213.

[34] VAN DIS E A M, BOLLEN J, ZUIDEMA W, et al. ChatGPT: five priorities for research[J]. Nature, 2023, 614(7947): 224–226.

作者简介

**徐道柱**（1982– ），男，博士，西安测绘研究所副研究员，主要研究方向为地理信息处理与应用。

**赵凯琳**（1995– ），女，中国科学院计算技术研究所博士生，主要研究方向为小样本学习、知识抽取。

**康栋**（1991– ），男，航天恒星科技有限公司工程师，主要研究方向为知识图谱。

**马超**（1988– ），男，博士，西安测绘研究所助理研究员，主要研究方向为地理信息智能化处理。

**冯禹铭**（1999– ），男，中国科学院计算技术研究所硕士生，主要研究方向为命名实体识别、增量学习。

**李紫宣**（1995– ），男，博士，中国科学院计算技术研究所助理研究员，主要研究方向为知识图谱、自然语言处理。

**弋步荣**（1984– ），男，航天恒星科技有限公司工程师，主要研究方向为遥感应用、人工智能、知识图谱。

**靳小龙**（1976– ），男，博士，中国科学院计算技术研究所研究员、博士生导师、CCF高级会员，主要研究方向为大数据知识工程、知识图谱。

**收稿日期：** 2023-08-03
**通信作者：** jinxiaolong@ict.ac.cn