doi:10.11920/xnmdzk.2025.03.009

基于元学习的中文史料少样本命名实体识别研究

陈 涛,殷 锋,郭伟超

(西南民族大学计算机与人工智能学院,四川 成都 610041)

摘 要:命名实体识别通过识别史料中的人名、地名、事件等实体,能够构建结构化知识库,促进历史信息的语义关联,助力历史事件的重构与分析.由于中文史料中标注语料稀缺,少样本问题较为突出.对此,在自适应裕度元学习三元组网络框架中集成数据增强技术,并使用针对中文的基于双向编码器表示的稳健优化的双向编码器表征预训练模型(robustly optimized bidirectional encoder representations from transformers pretraining approach for Chinese,Chinese-RoBERTa)优化中文长文本语义表征能力,提出融合数据增强的自适应裕度三元组网络命名实体识别方法.实验表明,本方法在少样本场景下表现显著,在 5-way 1-shot 任务和 5-way 5-shot 任务中 F1 值分别达到 86.68%和 92.78%,验证了其在低资源中文史料场景下的有效性.提出的方法解决了中文史料命名实体识别标注数据稀缺的少样本问题,同时为更广泛的低资源历史文本信息抽取任务提供有益参考.

关键词:元学习;少样本;数据增强;命名实体识别

中图分类号:TP391.1 **文献标志码**:A **文章编号**:2095-4271(2025)03-0308-07

Research on named entity recognition of few samples of Chinese historical materials based on metalearning

CHEN Tao,YIN Feng,GUO Weichao

(School of Computer Science and Artificial Intelligence,Southwest Minzu University,Chengdu 610041,China)

Abstract:Named entity recognition (NER) enables the construction of structured knowledge bases by identifying entities such as person names,locations,and events in historical texts,thereby facilitating semantic associations of historical information and supporting the reconstruction and analysis of historical events.Due to the scarcity of annotated corpora in Chinese historical documents,the few-shot learning problem is particularly pronounced. To address this,data augmentation techniques were integrated into the adaptive margin meta-learning triplet network framework and the semantic representation capabilities were enhanced for Chinese long texts using the Chinese-RoBERTa (Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach for Chinese) model.This led to the proposal of an adaptive margin triplet network-based NER method that incorporated data augmentation.Experimental results demonstrated the effectiveness of this method in few-shot scenarios,achieving F1 scores of 86.68% and 92.78% in 5-way 1-shot and 5-way 5-shot tasks,respectively,validating its efficacy in low-resource Chinese historical text scenarios.The proposed approach addressed the challenge of scarce annotated data for NER in Chinese historical texts under few-shot conditions while providing valuable insights for broader low-resource historical text information extraction tasks.

Keywords:meta-learning;small sample size;data enhancement;named entity recognition

收稿日期:2025-03-01

通信作者:殷锋(1972-),男,教授,博士,研究方向:人工智能和数据挖掘等.E-mail:yf_eagle@swun.edu.cn

基金项目:四川省教育信息技术研究资助项目(DSJ2022036);西南民族大学中央高校基本科研业务费专项资金资助项目(ZYN2025109)

中文史料作为中华文明的重要载体,通过命名实体识别技术,能从海量的史料中准确识别和提取出史料中的地名、人名、事件等实体信息,对文化传承与历史研究具有关键意义[1-3].命名实体识别(named entity recognization,NER)于 1995 年 11 月在第六届 MUC 会议(MUC-6, the Sixth Message Understanding Conference)上被提出,是信息抽取和信息检索中一项重要的任务,现已广泛应用于各个领域[4].基于深度学习的命名实体识别方法依赖大规模的标注语料[5],而中文史料领域受限于中文语法复杂性、专业术语多样性,标注成本指数级增长,致使标注语料规模受限,严重制约模型在低资源场景下的性能.少样本命名实体识别旨在使用少量地标注数据,识别模型未曾遇到过的实体,能有效解决在少样本或数据标注稀缺时的性能问题.

目前少样本命名实体识别方法主要可分为四类.第一类是基于结构优化的方法[6-7],该方法主要是优化模型的结构与特征提取,能够在复杂或不完整的数据场景下提高模型的准确性和鲁棒性,但难以捕捉长距离语义依赖.第二类是基于模型微调与优化的方法[8-9],侧重微调现有模型,可以快速适应新任务,有效减少对大量标注数据的依赖,但在增量学习场景下存在灾难性遗忘.第三类是基于对比学习方法[10],通过引入对比学习,增强类别区分能力,提高分类准确率,减少错误分类,但该方法对伪标签噪声敏感.第四类是基于元学习的方法[11],此类方法通过元学习框架构造实体类型三元组,根据样本与三元组距离并结合语义信息进行实体类型预测,模型能够高效学习新类别的实体,提高模型的跨领域、跨任务的泛化能力,具备快速适应新任务的特点.其固定边距设计易引发原型空间混淆,且未能充分解决样本稀疏问题,在训练样本规模受限时,存在元任务构建过程采样困难的问题.

为突破上述瓶颈,研究提出融合数据增强的自适应裕度三元组网络命名实体识别方法(EDA-RoBERTa-MeTNet).充分解决了中文史料领域低资源集场景下的样本稀疏性问题,并优化中文长文本的语义表征能力,极大提升模型在低资源场景下的性能.该方法将数据增强和动态边距的三元组元学习框架相结合,为中文史料少样本 NER 研究提供新范式.

1　研究现状分析

元学习又被称为“学会学习”[12],是目前深度学习领域的重点研究方向之一.元学习提供了一个新的学习范式,先面向多个任务联合训练学习到有用的先验知识,然后在未来新场景任务时利用先验知识引导训练过程更快更好,增强学习器在多任务时的泛化能力,是一种通过运用在少样本任务中学到的知识使得模型快速适应新任务的方法.目前常用于命名实体识别任务的元学习方法主要分为基于优化和基于度量两大类.

基于优化的少样本命名实体识别方法的优点是与模型结构无关,并且能显著减少对新任务的数据依赖,其代表方法是基于模型不可知元学习(Model-Agnostic Meta-Learning, MAML)[13]的命名实体识别方法.模型不可知元学习方法提供了一种通用的方法来适应不同领域的参数,其本质是通过学习一个好的参数初始化来解决少样本学习问题,找到一个好的模型初始参数,使得模型能够通过少量梯度更新快速适应新任务.

基于度量的少样本命名实体识别方法优点则是相似性度量结果能直观反映样本间的语义关联,且学习的度量空间可迁移到新类别或新任务.其代表方法是基于原型网络(Prototypical Networks, ProtoNet)[14-16]的少样本命名实体识别方法.此类方法通过构建有效的度量学习机制学习任务之间的相似性,使模型能在新任务上进行快速识别和分类.

文献[13]指出 MAML 基于优化的元学习机制对任务分布的高度敏感性.即需要训练任务和测试任务来自相似的分布,如果训练任务和测试任务在本质上有较大差异,模型的性能可能会显著下降.跨领域迁移时,任务参数空间的几何结构差异会导致二阶梯度更新的方向偏移,使得模型难以找到具有跨任务适应性的初始化参数.这种现象在低资源场景下尤为明显,因为有限的训练样本无法支撑梯度下降路径的有效校正.Han 等人[16]基于原型网络提出具有自适应裕度的元学习三元组网络(MeTNet),引入动态可调的类间距离约束机制,改进原型网络的实体类型推理方式,有效解决原型网络无法处理类内数据分布复杂的问题.但该方法在训练数据规模受限时,模型在元学

习阶段构建支持集与查询集的过程中,会因样本多样性不足而陷入参数空间的局部最优解,导致元任务构建过程中停滞的问题;该方法在处理长文本序列时,上下文依赖关系的复杂性加剧了样本稀疏性对模型性能的制约.

2 融合数据增强的自适应裕度三元组网络命名实体识别方法

使用数据增强技术和使用 Chinese-RoBERTa 模型作为文本编码器改进具有自适应裕度的元学习三元组网络的命名实体识别方法,提出融合数据增强的自适应裕度三元组网络命名实体识别方法(EDA-RoBERTa-MeTNet).先将原始数据通过数据增强生成更多样本,解决构建支持集和查询集时,样本多样性不足导致的元任务构建停滞问题.通过 Chinese-RoBERTa 模型进行文本编码,增强中文长文本的语义表征能力,解决长文本中实体识别效果不佳的问题.使用多层感知机(MLP)构建三元组网络,用于学习支持集和查询集之间的关系.使用 MeTNet 模型中的改进的三元组损失函数计算查询集和支持集的损失,模型根据反馈更新参数以最小化两者之间的距离.最终,模型根据优化后的参数进行分类预测,并输出结果.模型结构如图 1 所示,图中红色线框标识了框架的核心改进模块,红色虚线框为文本数据增强模块,红色实线框为文本编码模块.

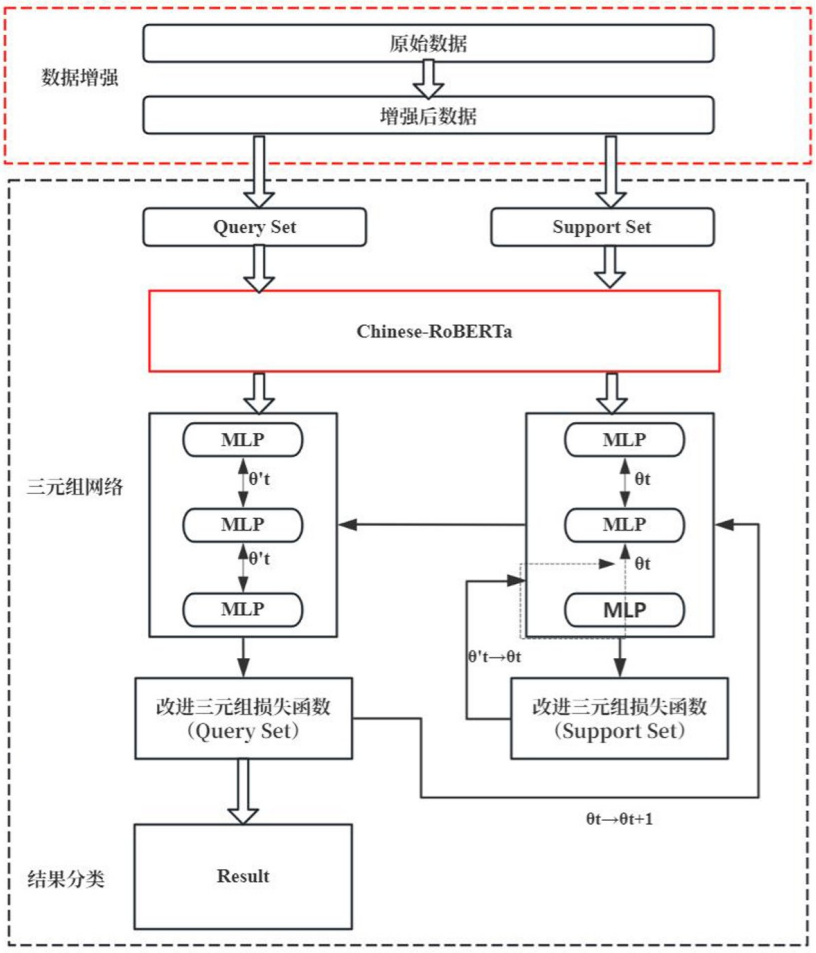

图 1 EDA-RoBERTa-MeTNet 模型结构

Fig.1 EDA-RoBERTa-MeTNet model structure

2.1 文本数据增强

基于简单数据增强(Easy Data Augmentation, EDA)[17]针对史料数据进行数据增强,在保证文本数据语义完整的前提下进行多维增强,本次数据增强主要分为四个维度:随机删除字符、非实体词同义替换、语料裁减以及语料拼接[18].具体的增强方法如表 1 所示.

表 1　文本数据增强方法

Table 1　Text data enhancement methods

级别	增强方法	描述
字符	RandomCharDelete	在给定单词中随机删除字符
单词	WordEmbsSubstitute	使用词嵌入模型 Word2Vec 预测的同义词替换文本中选定的词
句子	RandomSentClip	随机裁剪选定目标句
句子	RandomSentSplice	将选定目标句与其任意相邻句子拼接

通过非实体同义词替换可以在不破坏语料语义的情况下对数据量进行扩充,实体词同类替换操作保留了句子整体语义完整性的同时丰富了语料中命名实体场景的多样性.制定了对语料的裁减规则,首先以逗号为标识符进行裁减,其次为保证语料句法上的完整性,裁减后的句子不能只包含实体.裁减后的语料构成了后续语料拼接的语料库,对部分存在逻辑规则的语料进行拼接,丰富文本的上下文内容.

Word2Vec 是一种从文本语料中无监督学习语义知识,利用神经网络模型生成词嵌入的模型,旨在把语义上相近的词投射到向量空间中的近似位置,以便语义相似的词在向量空间中距离较近,它也可以用来获取指定单词的同义词.Word2Vec 包含 CBOW 和 Skip-gram 两种框架的模型,研究训练一个 CBOW 框架的 Word2Vec 模型用于词嵌入数据增强,词向量维度设置为 100.

2.2　文本编码

成都地区史料文本数据是中文史料数据集,经分析后发现该数据集中数据的最大文本长度较长,选择使用最长输入序列更长的 Chinese-RoBERTa 模型作为文本编码器,将每个字表示在一个低维嵌入向量中.具体地说,给定 n 个字$[x_1, x_2, \ldots, x_n]$的序列,将 Chinese-RoBERTa 模型中的最终隐藏层的输出作为x_i的初始表示h_i如式(1)所示.

$$[h_1, h_2, \ldots, h_n] = \text{RoBERTa}\varphi([x_1, x_2, \ldots, x_n]). \quad (1)$$

在 BERT[19] 中,遮掩是在预处理阶段进行的,每个句子在整个训练过程中保持相同的遮掩模式.Chinese-RoBERTa 模型则是在每个训练步骤中动态生成遮掩模式,使模型在不同的训练步骤中看到更多的上下文变体,从而提高模型的鲁棒性.并且 Chinese-RoBERTa 模型使用了更大的批次和更长的序列长度,使得模型可以在更长的上下文中进行学习,从而提升对长文本的理解能力.

3　实验

在少样本命名实体识别任务中,N-way K-shot 表示在训练和评估时,模型需要处理 N 种不同的实体类别(ways),并且每个类别只有 K 个示例(shot)来进行学习.实验主要分析不同模型 5-way 1-shot 任务和 5-way 5-shot 任务的性能.为验证 EDA-RoBERTa-MeTNet 方法的有效性,从域内对比研究和消融研究设计实验.

3.1　实验设置

1)史料文本数据集

数据集是命名实体识别的关键部分,它决定了在数据集上训练的模型是否适用于实际问题.目前没有成都地区史料这一特定领域的公共数据集,创建了一个自定义成都地区史料命名实体识别数据集.共收集到成都地区年鉴部分史料 253 本.使用“BIO”标签方法对分词后的文本进行单字序列标注.部分语料处理结果如表 2 所示.

表 2　部分史料语料库标注结果样例

Table 2　Partial examples of annotation results in the historical data corpus

序号	词语	标记
1	来	O
2	自	O
3	石	B-地区
4	羊	I-地区
5	街	I-地区
6	道	I-地区
7	的	O
8	3	O
9	0	O
10	0	O
11	余	O
12	名	O
13	民	O
14	间	O
15	艺	O
16	人	O

共标注了 10 类实体,包括地区、时间、人口、作品、人物、组织、面积、美食、民俗、节日活动,其中数据集共计 14 220 条数据,包含 40 000 多个实体.最后将

数据集 80%作为训练集,10%为验证集,10%为测试集,具体数量如表 3 所示.

表 3 实体标注数量

Table 3 Entity annotation counts

实体类别	实体数量/个
地区	10 271
时间	5110
人口	10 285
作品	2 150
人物	980
组织	1 825
面积	1 858
美食	2 449
民俗	1 201
节日活动	4 065

2)评价指标

使用评价指标精确率(P)、召回率(R)和 F1 值来检验和评测模型效果.高精确率表示模型在预测正类时的准确性较高,能减少误报;高召回率表示模型能捕捉到大多数真实的正类样本;F1 值同时考虑了精确率和召回率,是综合性指标.具体计算公式如式(2)~式(4)所示.

$$P=\frac{T_p}{T_p+F_p}. \tag{2}$$

$$R=\frac{T_p}{T_p+F_n}. \tag{3}$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall}. \tag{4}$$

3)实验参数

实验使用基于优化方法的 MAML 模型[13]与基于度量方法的 Proto 模型[14]和 MeTNet 模型[16]作为基线模型.实验参数设置:最大输入文本 512,RoBERTa 初始学习率为 0.000 02,模型中其他参数的学习率为 0.000 1.实验具体参数如表 4 所示.

表 4 实验参数设置

Table 4 Experimental parameter settings

实验参数	参数值
meta_lr	0.000 1
train_iter	6 000
test_iter	500
val_iter	100
Batch_size	1
max_length	512

3.2 实验结果与分析

1)域内对比研究

域内对比研究是在自定义成都地区史料文本数据集上与三个基线模型进行对比实验.实验结果如表 5 和图 2 所示.

表 5 不同模型域内对比实验结果

Table 5 Experimental results of different models

模型	5-way 1-shot 任务			5-way 5-shot 任务		
	P/%	R/%	F_1/%	P/%	R/%	F_1/%
MAML	68.42	82.77	74.91	78.38	89.91	83.75
Proto	28.80	81.36	42.54	60.74	91.23	72.93
MeTNet	82.92	72.92	77.60	85.24	78.39	81.67
EDA-RoBERTa-MeTNet	90.57	83.11	86.68	94.57	91.06	92.78

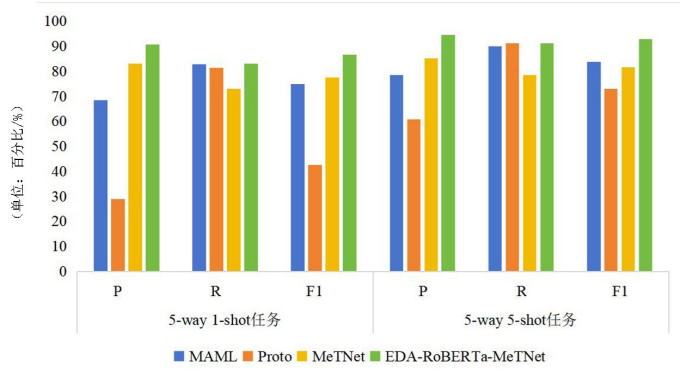

图 2 不同模型域内对比实验结果

Fig.2 Experimental results of different models

从域内对比实验结果可以看出,EDA-RoBERTa-MeTNet 在 5-way 1-shot 任务和 5-way 5-shot 任务的准确率和 F1 得分优于所有基线模型和 MeTNet 模型.与 MAML 相比,5-way 1-shot 任务准确率提升了 22.15%,F1 得分提升 11.77%;5-way 5-shot 任务准确率提升了 16.19%,F1 得分提升 9.03%.与 MeTNet 模型相比,5-way 1-shot 任务的准确率提升 7.65%,F1 得分提升 9.08%;5-way 5-shot 任务准确率提升 5.73%,F1 得分提升 4.46%.

2)消融研究

消融研究主要验证 EDA-RoBERTa-MeTNet 方法中数据增强和 Chinese-RoBERTa 模型的有效性.在成都地区史料文本数据集上设计了以下 4 个模型的对比实验.MeTNet 方法未使用数据增强和 Chinese-RoBERTa 模型;EDA-MeTNet 方法使用了数据增强;RoBERTa-MeTNet 模型使用 Chinese-RoBERTa 模型作为文本编码器但未使用数据增强方法.消融实验结果如表 6 和图 3 所示.

表 6　模型消融研究实验结果

Table 6　Experimental results of model ablation studies

模型	5-way 1-shot 任务			5-way 5-shot 任务		
	P/%	R/%	F_1/%	P/%	R/%	F_1/%
MeTNe	82.92	72.92	77.60	85.24	78.39	81.67
EDA-MeTNet	88.39	79.88	83.92	93.91	91.23	90.16
RoBERTa-MeTNet	86.35	69.27	76.87	88.84	86.70	88.32
EDA-RoBERTa-MeTNet	90.57	83.11	86.68	94.57	91.06	92.78

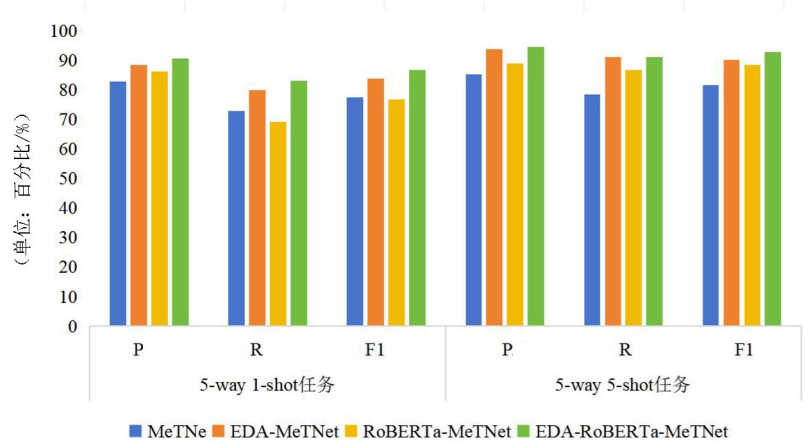

图 3　模型消融研究实验结果

Fig.3　Experimental results of model ablation studies

从消融实验结果可以看出,在 5-way 1-shot 任务和 5-way 5-shot 任务中,使用 Chinese-RoBERTa 模型作为文本编码器且使用数据增强的 EDA-RoBERTa-MeTNet 模型在准确率和 F1 得分都优于其他模型.对比 MeTNet 和 EDA-MeTNet 的实验结果,使用数据增强的 EDA-MeTNet 模型比未使用数据增强的 MeTNet 模型的准确率和 F1 得分都取得较大提升,可以证明数据增强对模型性能提升的有效性.对比 MeTNet 和 RoBERTa-MeTNet 的实验结果,在 5-way 1-shot 任务中,准确率有较大的提升,但召回率降低较大导致 F1 得分有略微的降低;在 5-way 5-shot 任务中,使用 Chinese-RoBERTa 模型作为文本编码器的模型的准确率和 F1 得分均有较大提升.

4　总结

本研究针对经典元学习 MeTNet 算法进行改进,提出融合数据增强的 EDA-RoBERTa-MeTNet 算法.该方法通过融合四种文本数据增强策略对原始数据进行多维增强,有效提升元任务构建过程中样本的多样性和分布合理性.同时,引入 Chinese-RoBERTa 模型作为深度语义编码器,通过其多层注意力机制强化对长文本中实体边界指示符的捕捉能力,进而改善复杂语

境下的实体识别鲁棒性.在成都地区史料文本数据集上,本研究提出的EDA-RoBERTa-MeTNet方法性能优于基于经典元学习的少样本命名实体识别方法,并且数据增强与Chinese-RoBERTa模型在提升模型性能方面均取得了良好效果,验证了文本数据增强和Chinese-RoBERTa模型的有效性.本研究使用的文本数据增强方法策略较少,也导致模型存在一些限制,随着支持集和查询集中每个类别中样本数量的增加,模型性能提升效果随之降低.在未来工作中,计划改进数据增强策略方法,并进一步提高模型识别效果.

参考文献

[1]李莉,宋涵,刘培鹤,等.基于数据增强和残差网络的敏感信息命名实体识别[J/OL].计算机应用,2024:1-7.[2025-02-26](2024-11-20).https://kns.cnki.net/kcms/detail/51.1307.TP.20241119.1549.012.html.

[2]项恒,杨明友,李猛.基于BiLSTM-CRF的航行通告命名实体识别研究[J].计算机科学,2024,51(S2):125-130.

[3]潘正高.基于规则和统计相结合的中文命名实体识别研究[J].情报科学,2012,30(5):708-712+786.

[4]杨欣怡,何小海,滕奇志,等.基于语义联合的跨模态命名实体识别[J/OL].计算机工程:1-9[2025-03-26].https://doi.org/10.19678/J.issn.1000-3428.000069927.

[5]郭云飞,温雪岩,焦燕,等.融合多特征与半监督学习的命名实体识别研究[J/OL].现代电子技术,2024:1-9.[2025-02-26](2024-10-15).https://kns.cnki.net/kcms/detail/61.1224.Tn.20241015.1602.002.html.

[6]江汀莹,线岩团,王红斌.结合近邻分析的小样本命名实体识别方法[J].现代电子技术,2023,46(19):88-94.

[7]戚荣志,周俊宇,李水艳,等.基于细粒度原型网络的小样本命名实体识别方法[J].软件学报,2024,35(10):4751-4765.

[8]吕海啸,李益红,周晓谊.前缀调优的少样本命名实体识别[J].计算机科学与探索,2024,18(8):2180-2189.

[9]吕明翰,黄琪,罗文兵,等.基于标签提示和门控模块的少样本命名实体识别[J].中文信息学报,2024,38(9):117-125.

[10]陈妍,辛逍,肖晓丹.基于提示学习的小样本命名实体识别[J].现代计算机,2024,30(17):49-54.

[11]张越,王长征,苏雪峰,等.基于标签语义信息感知的少样本命名实体识别方法[J].北京大学学报(自然科学版),2024,60(3):413-421.

[12]THRUN S,PRATT l.Learning to learn:Introduction and overview[M]//Learning to learn.Boston,MA:Springer US,1998:3-17.

[13]FINN C,ABBEEL P,LEVINE s.Model-agnostic meta-learning for fast adaptation of deep networks[EB/OL].2017:1703.03400.https://arxiv.org/abs/1703.03400v3.

[14]Snell J,Swersky K,Zemel R.Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems.Red Hook,NY,USA:Curran Associates InC.,2017:4080-4090.

[15]KUMAR R,GOYAL S,VERMA A,et al.ProtoNER:Few shot incremental learning for named entity recognition using prototypical networks[C]//Business Process Management Workshops.Cham:Springer Nature Switzerland,2024:70-82.

[16]HAN C C,ZHU R Y,KUANG J,et al.Meta-learning triplet network with adaptive margins for few-shot named entity recognition[J].CoRR,2023,abs/2302.7739.

[17]WEI J,ZOU K.EDA:easy data augmentation techniques for boosting performance on text classification tasks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP).Hong Kong,China:Association for Computational Linguistics,2019:6382-6388.

[18]林娜,岳希,唐聃.基于数据增强和损失平衡的机电领域命名实体识别[J/OL].计算机工程与应用,2024:1-12.[2025-02-26](2024-04-23).http://kns.cnki.net/kcms/detail/11.2127.TP.20240423.1216.004.html.

[19]DEVLIN J,CHANG M W,LEE K,et al.BERT:pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the NortH.Minneapolis,Minnesota:Association for Computational Linguistics,2019:4171-4186.

(责任编辑:张阳,付强,和力新,肖丽;英文编辑:周序林,郑玉才)