软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2024,35(3):1107−1124 [doi: 10.13328/j.cnki.jos.007069]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 融合多模态数据的小样本命名实体识别方法*

张天明, 张 杉, 刘 曦, 曹 斌, 范 菁

(浙江工业大学 计算机科学与技术学院, 浙江 杭州 310023)

通信作者: 曹斌, E-mail: bincao@zjut.edu.cn

**摘 要**: 作为自然语言处理领域的关键子任务, 命名实体识别通过提取文本中的关键信息, 帮助机器翻译、文本生成、知识图谱构建以及多模态数据融合等许多下游任务深度理解文本蕴含的复杂语义信息, 有效地完成任务. 在实际生活中, 由于时间和人力等成本问题, 命名实体识别任务常常受限于标注样本的稀缺. 尽管基于文本的小样本命名实体识别方法已取得较好的泛化表现, 但由于样本量有限, 使得模型能提取的语义信息也十分受限, 进而导致模型预测效果依然不佳. 针对标注样本稀缺给基于文本的小样本命名实体识别方法带来的挑战, 提出了一种融合多模态数据的小样本命名实体识别模型, 借助多模态数据提供额外语义信息, 帮助模型提升预测效果, 进而可以有效提升多模态数据融合、建模效果. 该方法将图像信息转化为文本信息作为辅助模态信息, 有效地解决了由文本与图像蕴含语义信息粒度不一致导致的模态对齐效果不佳的问题. 为了有效地考虑实体识别中的标签依赖关系, 使用 CRF 框架并使用最先进的元学习方法分别作为发射模块和转移模块. 为了缓解辅助模态中的噪声样本对模型的负面影响, 提出一种基于元学习的通用去噪网络. 该去噪网络在数据量十分有限的情况下, 依然可以有效地评估辅助模态中不同样本的差异性以及衡量样本对模型的有益程度. 最后, 在真实的单模态和多模态数据集上进行了大量的实验. 实验结果验证了该方法的预测 *F*1 值比基准方法至少提升了 10%, 并具有良好的泛化性.

**关键词**: 命名实体识别; 多模态数据; 小样本学习; 元学习; 去噪网络

**中图法分类号**: TP18

中文引用格式: 张天明, 张杉, 刘曦, 曹斌, 范菁. 融合多模态数据的小样本命名实体识别方法. 软件学报, 2024, 35(3): 1107−1124. http://www.jos.org.cn/1000-9825/7069.htm

英文引用格式: Zhang TM, Zhang S, Liu X, Cao B, Fan J. Multimodal Data fusion for Few-shot Named Entity Recognition Method. Ruan Jian Xue Bao/Journal of Software, 2024, 35(3): 1107−1124 (in Chinese). http://www.jos.org.cn/1000-9825/7069.htm

## Multimodal Data Fusion for Few-shot Named Entity Recognition Method

ZHANG Tian-Ming, ZHANG Shan, LIU Xi, CAO Bin, FAN Jing

(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

**Abstract**: As a crucial subtask in natural language processing (NLP), named entity recognition (NER) aims to extract the import information from text, which can help many downstream tasks such as machine translation, text generation, knowledge graph construction, and multi-modal data fusion to deeply understand the complex semantic information of the text and effectively complete these tasks. In practice, due to time and labor costs, NER suffers from annotated data scarcity, known as few-shot NER. Although few-shot NER methods based on text have achieved sound generalization performance, the semantic information that the model can extract is still limited due to the few samples, which leads to the poor prediction effect of the model. To this end, this study proposes a few-shot NER model based on the multi-modal dataset fusion, which provides additional semantic information with multi-modal data for the first time, to help the model prediction and can further effectively improve the effect of multimodal data fusion and modeling. This method converts image

* 基金项目: 国家自然科学基金(62276233, 62302451); 浙江省自然科学基金(LQ22F020018); 浙江省重点研发项目(2023C01048)
本文由“面向多模态数据的新型数据库技术”专题特约编辑彭智勇教授、高云君教授、李国良教授、许建秋教授推荐.
收稿时间: 2023-07-15; 修改时间: 2023-09-05; 采用时间: 2023-10-24; jos 在线出版时间: 2023-11-08
CNKI 网络首发时间: 2023-12-22

information into text information as auxiliary modality information, which effectively solves the problem of poor modality alignment caused by the inconsistent granularity of semantic information contained in text and images. In order to effectively consider the label dependencies in few-shot NER, this study uses the CRF framework and introduces the state-of-the-art meta-learning methods as the emission module and the transition module, respectively. To alleviate the negative impact of noisy samples in the auxiliary modal samples, this study proposes a general denoising network based on the idea of meta-learning. The denoising network can measure the variability of the samples and evaluate the beneficial extent of each sample to the model. Finally, this study conducts extensive experiments on real unimodal and multimodal data sets. The experimental results show the outstanding generalization performance of the proposed method, where the proposed method outperforms the state-of-the-art methods by 10 *F*1 scores in the 1-shot setting.

**Key words**: named entity recognition; multi-modal data; few-shot learning; meta learning; denoising network

命名实体识别(named entity recognition, NER)任务旨在将文本中的命名实体, 例如时间、地点、人名等, 提取出来, 并分类到预定义的类别中[1], 作为自然语言处理领域中的关键子任务, 命名实体识别任务可以提取出文本中的关键词信息, 从而帮助模型更好地理解文本所蕴含的语义信息, 进而更加高效地完成各种下游任务, 比如信息抽取[2]、机器翻译[3]、文本生成[4]和知识图谱构建[5]等. 除此之外, 近两年更是在多模态数据建模中发挥了重要的作用[6,7], 帮助模型挖掘文本模态中的复杂语义信息, 进而有效地进行多模态融合, 实现多模态数据的高效利用.

近些年, 深度神经网络模型被广泛应用于命名实体识别任务上, 并达到了显著的效果. 然而, 深度模型的强大性能离不开海量训练数据集的支持. 在现实世界中, 由于标注数据需要耗费大量的人力和时间成本, 所以大规模的标注数据集往往难以获取. 无论是基于深度学习的主流命名实体识别方法[8−11], 还是基于统计学习的传统命名实体识别方法[12−14], 使用少量(小样本)标注数据训练模型都会造成模型过拟合, 进而导致命名实体识别效果不佳. 因此, 如何在标注数据十分有限的情况下保证模型的预测效果, 成为一个重要的挑战.

目前, 小样本命名实体识别(few-shot NER)方法主要分为以下 3 类.

(1) 数据增强[15]: 数据增强方法从扩充数据集规模的角度, 将小样本命名实体识别任务转化成一般命名实体识别任务. 具体地, 数据增强方法使用同义词替换、机器翻译以及添加微弱噪声等方法对已有的小样本数据集进行改动以产生新数据, 从而达到扩充训练集的目的.

(2) 模型预训练[16]: 使用海量开源数据集对模型进行预训练, 使得模型学习到更加适用于命名实体识别任务的通用语义空间. 此外, 还可以加入提示学习(prompt learning), 输入提示信息引导模型输出正确的预测值.

(3) 原型网络[17,18]: 旨在学习一个通用的语义空间, 在该空间中, 每个类别的实体聚成一个簇, 每个簇中心称为该类别的原型, 通常是该类别所有实体表示的平均值. 在进行预测时, 通过计算被预测实体与每个类别原型之间的相似度, 相似度最高的簇作为被预测实体的类别.

然而, 现有的小样本命名实体识别方法专注于从稀缺的文本中挖掘语义信息, 但文本中包含的有效语义信息是十分有限的, 进而导致模型泛化能力受限. 为了提升模型的泛化能力, 本文将多模态数据用于小样本命名实体识别任务, 以补充单模态文本数据缺失的有效语义信息. 多模态学习旨在借助其他模态的数据帮助模型有效地提升预测效果[19,20], 例如借助图像和语音等其他模态的数据帮助模型学习文本分类任务. 其中, 在大规模样本的条件下, 使用图像信息帮助提升命名实时, 识别任务的效果已被证明是有效的[21]. 因此, 本文考虑使用图像数据作为辅助模态数据, 来帮助模型学习面向文本的小样本命名实体识别(multimodal few-shot NER)任务.

现有的多模态命名实体识别方法指出, 文本信息与图像信息存在语义密度不一致的问题, 即文本包含的语义信息相对具体而图像包含的信息相对抽象, 文本与图像信息无法有效对齐, 进而影响多模态数据融合效果[21]. 为此, 本文引用 Wang 等人[22]提出的思想, 将图片信息转化成文本信息, 进而有效地将图片信息映射到文本语义空间, 使得两种模态信息能够有效融合. 此外, 为了获取有效的文本表示, 本文使用原型网络构建通用语义空间. 原型网络通过大量源领域样本学习通用的语义空间, 这使得不同领域之间的标签会相互影响, 造成误分类. 鉴于此, 本文引用最先进的基于原型网络的方法 L-TapNet[23]. L-TapNet 在进行具体任务时, 会将

实体映射到一个新的空间, 进而有效地避免了其他领域的干扰.

然而, 现有的多模态小样本方法专注于多模态数据的有效对齐以及通用表示空间的学习, 而忽略了辅助模态中的噪声数据对模型产生的负面影响. 例如: 在命名实体识别任务中, 图像数据常作为辅助模态数据提供额外的语义信息, 进而帮助模型提高文本分类效果, 但是多模态数据往往来源于大量的社交媒体平台, 这类数据存在很严重的图文不符的情况. 具体来说, 用户会使用图片加文本的方式表达观点, 而图片的内容可能与文本内容截然不同. 用户通过这种表达方式, 使自己的观点变得生动形象或讽刺意味浓厚. 图1是一则由用户发表在社交平台的动态. 文本的内容为“家里有只大老鼠”, 而配图是一只躺在床上的德牧犬. 用户旨在使用比喻的方式生动地表达出自己的观点. 由于文本与图片信息不符, 在进行实体识别时, 模型旨在将文本中的“老鼠”分类为“鼠科”, 但是由于图片表示“犬科”的信息, 可能对模型的识别产生负面的影响,进而造成误分类. 因此, 本文提出加入去噪网络, 弱化辅助模态中的噪声数据对模型的干扰.

尽管现有的基于小样本的方法使用超参数作为权重因子来衡量辅助模态数据对模型的影响程度, 但使用统一的超参数无法有效地衡量不同样本的差异性, 导致模型效果依然不佳. 此外, 传统的去噪网络依赖大规模数据集, 将其直接应用于小样本环境会导致模型过拟合. 因此, 本文提出了基于元学习的通用去噪网络, 在数据量十分受限的情况下, 能够有效地衡量样本间的差异性并评估其对模型的有益程度, 进而提高模型预测效果. 元学习旨在从丰富源领域学习先验知识, 然后利用先验知识指导模型在少量样本的情况下学习目标领域的任务. 基于上述思想, 本文设计了去噪网络, 将单模态文本表示和考虑了图像信息的双模态文本表示拼接作为输入, 输出是单模态文本表示和双模态文本表示的权重因子. 其中, 去噪网络参数作为先验知识, 通过大量源领域数据进行学习, 使其在小样本的情况下, 依然可以有效地评估辅助模态对模型的影响程度.


图 1 噪声模态样本示例

本文工作的主要贡献可以总结为以下 3 点.

(1) 提出一种多模态小样本命名实体识别模型 MFNER. 该模型借助其他模态数据(图像数据)信息,在小样本的情况下, 能够更好地理解文本语义信息, 进而提高模型的预测效果.

(2) 提出了基于元学习思想的通用去噪网络. 该网络使用元学习思想, 将网络参数作为元参数进行学习, 进而得到通用的去噪网络, 使其在样本量十分受限的情况下, 有效地衡量不同辅助模态样本的差异性并评估出不同样本对模型的有益程度, 帮助模型提升预测效果.

(3) 在多个真实的单模态和多模态 NER 数据集上进行了大量的实验评估, 验证了 MFNER 与基准方法相比在 $F1$ 值上至少提升了 10%, 且具有良好的泛化性.

本文第 1 节阐述多模态小样本实体识别的相关工作. 第 2 节介绍小样本实体识别 few-shot NER 的基本概念. 第 3 节详细阐述多模态小样本实体识别模型 MFNER. 第 4 节通过详尽的实验评估分析 MFNER 的预测性能. 第 5 节总结全文.

## 1 相关工作

本节介绍多模态小样本命名实体识别的相关工作. 第 1.1 节回顾多模态命名实体识别的相关工作, 包括考虑语音模态、字形模态和图像模态的多模态命名实体识别方法. 第 1.2 节介绍多模态小样本的相关工作, 主要分为基于度量学习和基于模型预训练的多模态小样本两类研究工作.

### 1.1 多模态命名实体识别

近年来, 多种模态的数据被广泛用于提高命名实体识别效果. 根据不同的辅助模态, 多模态命名实体识别方法可以分为以下 3 类.

(1) 文本+语音. 中文文本中词语之间没有间隔, 导致实体的边界难以确认. 由于人们在说话时会在词语之间作适当的停顿, 因此, 语音模态被作为辅助模态帮助模型确认文本中实体的边界. Sui 等人[24]使用 CNN 下采样的梅尔滤波器组特征作为语音特征表示, 与使用 BERT 获取的文本表示融合进行实体识别.

(2) 文本+字形结构. 汉字的特殊字形结构也常用于辅助实体识别. Meng 等人[25]使用隶书、繁体字等古汉字的文字图片作为汉字结构信息, 并提出“田字格 CNN”对汉字图片的特征进行提取作为辅助信息. 他们把文本中的每个汉字拆解成部首, 然后使用 CNN 提取汉字的部首特征. Wu 等人[26]将汉字分解成部首作为辅助模态, 并用 CNN 提取部首特征, 然后使用 two-stream Cross-Transformer 与文本的嵌入进行融合.

(3) 文本+图片. 由于社交媒体文本存在语法错误、噪音多以及信息缺失等问题, 部分工作使用相关的配图提供有效的实体信息帮助模型预测. Moon 等人[27]首次提出了多模态命名实体识别任务(MNER), 使用 Bi-LSTM 和 CNN 分别获取文本嵌入和图像嵌入, 将其进行融合然后预测, 并提出了一个 MNER 数据集 SnapCaption. Sun 等人[28]提出了一种基于关系推断和视觉注意力的新型预训练多模态 NER 模型 RIVA, 通过门控制基于注意力的视觉线索, 关注图像对文本语义的作用. Wang 等人[22]认为, 文本模态特征和图像模态特征由不同的特征提取器获取, 无法很好对齐两种特征, 因此提出将图像信息转化为文本信息作为辅助模态,进而有效与文本特征进行融合.

值得指出的是: 上述所有的工作都依赖于大规模训练数据集, 在数据量受限的情况下会导致模型过拟合, 因此并不能有效地应用于小样本命名实体识别任务.

### 1.2 多模态小样本学习

由于基于度量学习和基于模型预训练的方法在小样本学习中的效果显著, 因此, 现有的多模态小样本学习方法大多也基于上述两种思想开展研究, 其目标专注于在上述两种框架下, 如何有效进行多模态数据融合. 在基于度量学习的多模态小样本工作中, Pahde 等人[29]使用对抗学习实现图像嵌入和文本嵌入的有效对齐, 通过对两种嵌入加权求和得到融合多模态信息的 cross-modal 原型, 进而达到利用文本信息辅助小样本模型进行图像分类的目的. Memmesheimer 等人[30]考虑将信号作为辅助模态嵌入到图像表示空间中, 并使用 CNN 模型提取多模态特征, 帮助小样本模型进行动作识别. Aktukmak 等人[31]提出了 ALMO, 一个面向任意多模态数据的小样本学习的框架, 使用特定编码器, 将样本与类别的属性数据从高维空间嵌入到一个通用的随机潜在空间中, 并使用非参分类器进行分类. 在基于模型预训练的多模态小样本工作[32]中, Tsimpoukelli 等人[33]提出了 Frozen 通用语言模型, 在小样本环境下完成许多下游任务, 例如图像分类和视觉问答. Frozen 先使用大规模文本数据预训练模型, 然后加入图像模型联合训练以实现多模态嵌入有效对齐. 在联合训练过程中, 不更新文本编码器参数. 为了解决社会关系严重不平衡的问题, Wan 等人[34]对电视剧和名著进行注释, 得到除文本和图像外的第 3 种模态数据, 并提出了 FL-MSRE 模型, 在样本量受限的情况下有效提取社会关系. Alayrac 等人[35]提出了视觉语言模型 Flamingo, 将视觉和文本单模态小样本模型有效链接起来. Flamingo 可以处理任意交错的视觉和文本数据序列, 以及不间断提取图像或视频作为输入, 进而有效地学习多模态上下文信息. 然而,现有的多模态小样本方法都忽略了辅助模态中噪声样本对模型预测的负面影响. 即使有工作[29]考虑了辅助模态的影响程度, 也只是使用简单的策略, 通过设置简单的超参数作为影响因子, 这样的方法并不能有效地评估不同噪声样本之间的差异性.

## 2 基础知识

本节介绍小样本命名实体识别的问题定义及相关基本概念. 首先, 第 2.1 节介绍问题定义; 而后, 第 2.2 节介绍应用于命名实体识别任务的条件随机场模型(CRF); 最后, 第 2.3 节介绍元学习(meta-learning)思想. 表 1 总结了本文中常用的符号.

表 1 常用符号

| 符号 | 描述 |
|---|---|
| $x=(x_1,x_2,..,x_L)$ | 一个含有 $L$ 个字符的词序列 |
| $y=(y_1,y_2,..,y_L)$ | 词序列 $x$ 对应的标签序列 |
| $T=\{T_1,T_2,..,T_m\}$ | 一个含有 $m$ 个源领域的集合 |
| $T'=\{T'_1,T'_2,...,T'_n\}$ | 一个含有 $n$ 个目标领域的集合 |
| $S$ | 支持集, 用于元学习模型中内层模型的训练集 |
| $Q$ | 查询集, 用于元学习模型中外层模型的训练集 |
| $K$ | 小样本数据集中每个标签对应的样本个数 |
| $F(y,x)$ | CRF 模型中给定 $x$ 得到标签序列 $y$ 的总分数 |
| $f_E$ | CRF 模型中给定 $x$ 得到标签序列 $y$ 的发射分数 |
| $f_T$ | CRF 模型中给定 $x$ 得到标签序列 $y$ 的转移分数 |
| $L$ | CRF 模型的 Loss 损失 |
| $\theta$ | 元学习模型中内层模型的参数 |
| $\Phi$ | 元学习模型中外层模型的参数 |
| $\alpha$ | 去噪网络中文本信息的有益程度 |
| $\beta$ | 去噪网络中图像信息的有益程度 |
| $E_t$ | 单模态文本嵌入 |
| $E_m$ | 考虑了图像信息的双模态文本嵌入 |
| $E_f$ | 融合了 $E_t$ 和 $E_m$ 的融合多模态嵌入 |

### 2.1 小样本命名实体识别

在训练集样本数量极少的情况下, 给定输入句子 $x=(x_1,x_2,..,x_L)$, 输出该句子对应的标签序列 $y=(y_1,y_2,..,y_L)$. 其中, $L$ 为句子长度. 通常, 一个领域定义为一个任务, 每个领域是由多个样本对$(x,y)$组成的集合. 在小样本命名实体识别任务中, 模型首先在源领域集合 $T=\{T_1,T_2,..,T_m\}$上训练以学习先验知识. 在先验知识的指导下, 模型在目标领域集合 $T'=\{T'_1,T'_2,...,T'_n\}$ 上进行测试. 其中, $T'$只包含极少数的标注样本 $S$, 我们将其称为支持集. $S$ 通常遵循 $K$-shot 原则, 即每个标签只有 $K$ 个样本. 基于上述描述, 我们给出小样本命名实体识别的定义. 定义如下: 给定一条查询样本 $x$ 和一个 $K$-shot 支持集 $S$, 小样本命名实体识别旨在找到 $x$ 的最优标签序列 $y^*$.

$$y^* = \arg\max_y P(y \mid x, S) \tag{1}$$

### 2.2 命名实体识别任务中的CRF框架

命名实体识别常被视为序列标注任务[36]. 作为经典的序列标注模型, 条件随机场(conditional random field, CRF)被广泛应用于命名实体识别任务. 在给定查询样本的情况下, 条件随机场模型通过考虑实体和标签之间的联系以及标签之间的依赖关系输出概率最大的标签序列. 因此, CRF 模型主要分为两个模块: 发射模块和转移模块. 其中, 发射模块用于计算实体被分类到标签的分数并将其称为发射分数, 转移模块用于计算标签之间的依赖转移分数. 以下给出条件随机场的一般化定义.

一般地, 条件随机场模型将候选标签序列的概率定义为 $p(y|x)$.

$$P(y \mid x) = \frac{\exp(F(y,x))}{\sum_{y' \in Y} \exp(F(y',x))} \tag{2}$$

其中, $F(y,x)$计算了将标签序列 $y$ 分配给 $x$ 的分数. 形式化地:

$$F(y,x) = \sum_{l=1}^{l} \left( f_E(y_l,x) + f_T(y_{l-1},y_l) \right) \tag{3}$$

具体来说, $F(y,x)$计算了序列中所有位置上的 $y_l$ 被分配到 $x_l$ 的分数和. 位置 $l$ 的分数包括发射分数 $f_E(y_l,x)$和转移分数 $f_T(y_{l-1},y_l)$两部分. 对于发射分数, 通常, CRF 将 $x_l$ 和 $y_l$ 嵌入表示的相似度作为发射分数. 其中, $x_l$ 考虑了文本的上下文语义信息. 对于转移分数, CRF 构建了一个规模为 $N\times N$ 的转移矩阵来记录标签之间的转移分数, $N$ 为标签的个数. 图 2 给出了转移矩阵的示例, 转移矩阵的每一行表示上一个词对应的标签, 每一列表示当前词对应的标签. 矩阵的分数值表示在上一个词的标签确定的情况下, $x_l$ 被分配到不同标签的概率. 例如: 在上一个词标签是“B-人”时, 当前词被分配到“I-人”的概率是 0.76.

| | O | B-人名 | B-时间 | I-人名 | I-时间 |
|---|---|---|---|---|---|
| O | 0.50 | 0.47 | 0.45 | 0.05 | 0 |
| B-人名 | 0.4 | 0.13 | 0.28 | 0.76 | 0.21 |
| B-时间 | 0.6 | 0.29 | 0.16 | 0 | 0.72 |
| I-人名 | 0.67 | 0.16 | 0.72 | 0.76 | 0 |
| I-时间 | 0.59 | 0.71 | 0.19 | 0.04 | 0.81 |

图 2 CRF 转移矩阵示例

最后, 基于公式(2)和公式(3), 定义 CRF 的 *Loss* 损失计算为

$$L=\sum_{(x,y)\in D}-\log P(y\mid x) \tag{4}$$

其中, $D$ 为训练集. 此外, 基于训练模型, 维特比算法[37]通常被用于从候选标签序列中找到最优标签序列.

### 2.3 元学习

元学习方法被广泛应用于小样本命名实体识别任务. 元学习思想旨在“学习如何学习”. 具体地, 元学习的目标是从丰富的源领域学习先验知识, 然后利用该先验知识指导模型在目标领域达到最大的泛化性能. 因此, 模型的优化过程分为内外两层优化: 内层的模型参数定义为$\theta$, 外层的元参数定义为$\Phi$. 在内层优化时, 从任务 $T_i$ 中随机采样取出支持集 $S$, 在给定支持集 $S$ 的情况下, 元参数$\Phi$指导模型$\theta$学习使其可以适应任务 $T_i$. 其中, 支持集 $S$ 包括 $N$ 个标签, 每个标签有 $K$ 个有限样本. 具体地, 在基于 CRF 框架的小样本命名实体识别任务中, 通过最小化公式(4)的损失函数 $L_{T_i}^{S}$, 得到模型在任务 $T_i$ 上的最优参数$\theta^*$. 形式化地:

$$\theta^*=\arg\min_{\theta}L_{T_i}^{S}(\theta,\Phi) \tag{5}$$

在外层优化时, 从任务 $T_i$ 中随机采样取出新的样本集合: 查询集 $Q$. 该集合被用来评估模型适应所有任务的泛化能力, 即通过最小化损失 $L_{T_i}^{Q}$ 调整元参数$\Phi$, 进而使模型达到更好的泛化性能. 最后, 在完成所有任务的两层优化后, 得到最优元参数$\Phi^*$. 综上所述, 元学习的目标被定义为: 给定所有任务的查询集, 最小化损失函数得到最优元参数$\Phi^*$. 形式化地:

$$\Phi^*=\arg\min_{\Phi}L_{T_i}^{Q}(\theta^*(\Phi)) \tag{6}$$

## 3 本文方法

本节详细阐述本文所提出的多模态小样本命名实体识别模型 MFNER. 接下来, 本节将从以下 5 个小节进行详细描述, 即模型概述、多模态信息提取模块、发射模块、转移模块和去噪网络.

### 3.1 模型概述

MFNER 旨在使用图像数据作为辅助模态, 为模型提供更丰富的语义信息, 进而帮助小样本模型提升预测效果. 首先, 由于文本和图像蕴含语义信息存在语义密度不一致问题, 本文引用 ITA 的思想, 将图像转换成文本, 从而实现文本与图像信息的有效对齐. 此外, 作为序列标注问题, 除了专注于学习实体和标签之间的联系, 命名实体识别任务还受益于考虑标签依赖关系. 因此, 本文使用 CRF 框架来捕捉实体识别任务的标签依赖关系, 并引用最先进的小样本方法 L-TapNet 和 CDT 机制[23]分别计算发射分数和转移分数. 最后, 由于图像辅助模态数据中可能存在噪声样本, 对实体识别起到负面影响, 为了减少噪声样本对模型的干扰, 本文提出了基于元学习的去噪网络, 在样本量十分受限的情况下, 依然可以有效考虑不同样本间的差异性, 并衡量出其对模型预测的有益程度.

基于上述思想, 本文提出了 MFNER 模型. 如图 3 所示, MFNER 主要由 3 个模块组成.

(1) 多模态信息提取模块. 该模块对图像数据进行预处理, 将其转换成文本形式, 并作为辅助模态信息

输入发射模块.

(2) 发射模块. 在该模块中, 基于文本数据以及辅助模态数据, 模型学习实体与标签之间的联系, 然后计算并输出对应的发射分数.

(3) 转移模块. 转移模块用于捕捉标签之间的依赖关系, 然后输出对应的转移分数.

最后, 基于计算出的发射分数和转移分数之和, 将候选标签序列中概率最高的标签序列作为查询样本对应的标签序列并输出.

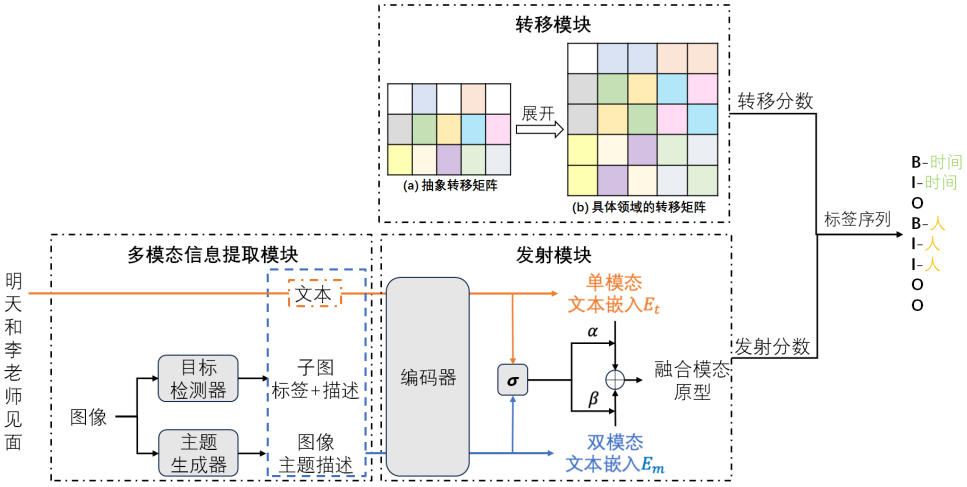



图 3 MFNER 架构图

下面对多模态信息提取模块、发射模块、转移模块以及去噪网络进行详细描述.

### 3.2 多模态信息提取

本文旨在使用图像数据作为辅助模态, 为小样本实体识别模型提供额外的语义信息, 从而弥补样本数量十分受限导致语义信息不足的问题, 进而提升模型的预测效果. 然而, 由于文本和图像数据表示的语义上下文信息存在语义密度不一致的问题, 导致两个模态的嵌入表示对齐效果不佳, 使得图像数据无法起到很好的辅助效果. 为此, 本文基于 ITA 方法的思想, 将图像数据转换为文本, 使得图像信息直接映射在文本语义空间, 从而避免不同模态嵌入表示无法有效对齐. 多模态信息提取模块对图像数据预处理, 将其转化为文本信息. 具体地, 图像信息按照粗细力度被分为以下两类.

(1) 全局信息. 将整个图像作为输入, 使用主题生成器生成整个图像的主题文本描述, 将其作为全局信息.

(2) 局部信息. 使用目标检测器, 将整个图像提取为多个子图, 并为所有子图生成对应的标签及描述, 然后将这些标签和描述作为局部信息.

最后, 将全局信息和局部信息拼接在一起. 基于上述预处理操作, 得到了转化为文本后的图像信息.

### 3.3 发射模块

在发射模块, 模型捕捉实体与标签之间的联系. 在小样本命名实体识别中, 原型网络被广泛应用. 原型网络的目标是为所有领域学习一个通用的语义空间 $E$, 在这个语义空间中, 同一标签对应的实体表示应该聚成一个原型簇[38]. 在本文中, 我们将语义空间 $E$ 中的实体表示记为 $E(x)$, 原型簇记为 $P$. 一般地, 原型是类别中所有的实体表示嵌入的平均值. 然而, 由于不同领域之间会相互干扰, 使得形成的原型簇存在偏差, 进而导致模型误分类[23]. 为此, 本文使用 L-TapNet 方法. L-TapNet 依然学习了一个通用的语义空间, 不同的是, L-TapNet 在目标领域进行预测时, 会将其映射在一个新的空间 $M$, 避免其他领域造成干扰. 最终, 查询样本的

嵌入表示 $E(x)$与标签 $c$ 的原型嵌入 $P_c$ 的相似度被作为发射分数 $f_E(c,x)$. 形式化地:

$$f_E(c,x)=Softmax\{SIM(M(E(x)),M(P_c))\} \tag{7}$$

其中, $SIM$ 为相似度计算方式. 由于 L-TapNet 只考虑了单模态数据的情况, 在本文中, 为了有效利用辅助模态信息, 需要进行多模态融合. 因此, 本文先将单模态文本数据和转化为文本的图像数据输入编码器, 分别得到单模态文本表示嵌入 $E_t$ 和考虑了图像信息的双模态文本嵌入 $E_m$. 再将上述两种表示嵌入进行融合, 得到融合模态表示嵌入 $E_f$. 为了避免辅助模态中的噪声数据对模型造成负面影响, 在对两种嵌入进行融合前, 本文将两种嵌入输入去噪网络$\sigma$, 分别得到两种嵌入的影响权重$\alpha$和$\beta$. 然后根据权重值对两种嵌入进行加权求和, 以保留有效的辅助信息. 形式化地:

$$E_f=\alpha\times E_t+\beta\times E_m \tag{8}$$

其中,

$$\alpha,\beta=\sigma(E_t,E_m) \tag{9}$$

关于去噪网络的具体细节, 将在第 3.5 节中进行阐述. 之后, 通过取平均的方式得到最终的融合模态原型 $P_c$, 在给定标签 $c$ 的 $m$ 个样本的情况下:

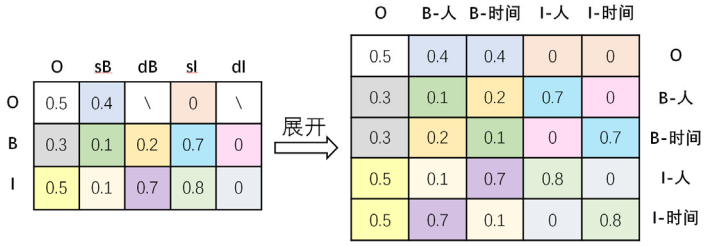
$$P_c=\frac{1}{m}\sum_{i=0}^{m}E_{f_i} \tag{10}$$

为了更好地理解发射模块, 本文使用图 3 中给出的例子进一步进行说明. 给定文本“明天和李老师见面”, 希望模型输出标签序列“B-时间、I-时间、O、B-人、I-人、I-人、O、O”. 因此, 在发射模块, MFNER 会计算出文本中每个字嵌入与所有标签(B-时间、I-时间、B-人、I-人和 O)的原型嵌入的相似度, 将其作为发射分数, 用于 CRF 总分数计算. 通过训练, 最终希望每个字与真实标签对应的相似分数尽量大.

### 3.4 转移模块

命名实体识别任务受益于考虑标签依赖转移. 然而, 直接将 CRF 应用于小样本命名实体识别任务会导致模型过拟合, 进而影响预测效果. 因此, 本文采用坍缩转移(collapsed dependency transfer, CDT)机制. CDT 利用元学习思想指导模型学习一个通用的抽象转移矩阵, 在目标领域中, 将抽象转移矩阵按照规则展开成适应具体领域的标签转移矩阵. 为了更好地理解 CDT 机制的原理, 本文结合图 3 中给出的例子展开详细的描述.

在元学习外层, CDT 构建抽象转移矩阵并进行学习. 如图 4(a)所示, 抽象转移矩阵为 3×5 的矩阵. 在构建过程中, CDT 将具体的标签依赖转移抽象成“O”“B”“I”到“O”“sB”“dB”“sI”“dI”之间的转移依赖. 其中, “dB”中的“d”表示两个具体标签的语义是不一致的. 例如, “B-人”到“B-时间”就属于抽象转移“B”到“dB”. 而“s”则表示语义一致.




(a) 抽象转移矩阵 (b) 具体领域的转移矩阵

图 4 抽象转移矩阵展开示例[23]

在元学习内层, CDT 将学好的通用抽象转移矩阵按照规则在具体领域中展开, 得到适应领域的转移矩阵. 由于图 3 中的标签集合为{B-时间、I-时间、B-人、I-人和 O}, 则将抽象转移矩阵展开为集合内 5 种具体标签之间的转移矩阵, 如图 4(b)所示. 展开规则为属于同一抽象标签转移的具体标签转移共享转移分数. 如图 4 所示: 在展开成右侧的具体任务转移矩阵时, “B-人”到“I-时间”和“B-时间”到“I-人”都属于抽象转移“B”到“dI”, 因此, 它们的转移分数都是 0.

在得到具体领域的转移矩阵后, MFNER根据矩阵和上个字的标签得到当前字被预测为不同标签的转移分数. 例如: 给定“李”的预测标签为“B-人”时, “老”被预测为“B-时间”“I-时间”“B-人”“I-人”和“O”的转移分数分别为0.2, 0, 0.1, 0.7和0.3.

最后, 基于由发射模块得到的发射分数和转移模块的转移分数, 根据公式(3), 计算出每个字被预测为每个标签的总分数, 分数最高的标签作为该词的预测标签.

### 3.5 去噪网络

由于图像中可能蕴含与实体相关的语义信息, 因此在本文中, 图像数据被作为辅助模态数据, 为有限的小样本文本数据提供额外的语义信息, 帮助模型提高泛化性能. 然而, 由于多模态数据中存在很严重的图文不符情况, 因此, 用于实体识别的图像数据存在大量的噪声样本, 这些样本不仅无法辅助模型完成预测, 甚至会对模型产生错误的引导, 进而导致模型误分类. 为此, 本文提出使用去噪网络, 缓解图像噪声样本对模型的干扰. 由于小样本实体识别任务中的样本数量十分受限, 使用传统的去噪方式会造成模型过拟合, 而目前基于超参数的小样本去噪方式又过于简单, 无法有效评估不同样本之间的差异性, 因此, 本文使用元学习思想, 在丰富的源领域学习通用的去噪网络, 使其在样本数量极少的目标领域, 依然可以有效评估样本之间差异性并衡量噪声样本对模型的有益程度. 值得注意的是: MFNER 只在发射模块加入了去噪网络, 而未在转移模块中使用. 由于发射模块需要计算多模态实体嵌入, 因此需要加入去噪网络评估图像模态信息的有效程度, 减少噪声影响, 进而有效融合文本嵌入和图像嵌入得到多模态实体嵌入. 而转移矩阵的值被作为模型的参数进行学习, 并没有直接用到实体嵌入, 不需要加入去噪网络.

去噪网络旨在对图像信息去噪, 以便得到更好的多模态融合表示. 多模态融合模态表示由以下两部分嵌入融合而成.

(1) 单模态文本嵌入. 相对多模态数据, 基于单模态数据的模型表现相对更加“稳定”. 因此, 单模态文本信息依然被作为输入的一部分用于提供可靠的语义信息.

(2) 考虑了图像信息的双模态文本嵌入. 将图像文本与单模态文本拼接输入 BERT 模型, 通过 BERT 的自注意力机制, 可以使得单模态文本有效学习到对自身有益的图像信息. 因此, 比起直接使用图像文本嵌入进行融合, MFNER 将考虑了图像信息的双模态文本嵌入作为融合嵌入的一部分, 有效地丢弃了无用的图像信息.

在此基础上, 去噪网络进一步对辅助模态信息进行去噪处理, 即分别为单模态文本嵌入 $E_t$ 和双模态文本嵌入 $E_m$ 生成影响因子$\alpha$和$\beta$(其中, $\alpha+\beta=1$), 旨在使用$\alpha$和$\beta$评估两个嵌入的有益程度. 具体地, 如果图像模态信息存在错误, 则希望去噪网络计算得到的$\beta$值偏小; 如果图像模态信息包含了额外的语义信息, 希望$\beta$值偏大,甚至可以大于$\alpha$值, 因为双模态文本嵌入可以提供更加丰富的语义信息. 通过学习去噪网络对不同的样本中不同的字生成合适的$\alpha$和$\beta$, MFNER 实现了根据样本内容模型自动判断并完成去噪任务的目的.

去噪网络模型架构如图 5 所示, 将每个字的 $E_t$ 和 $E_m$ 拼接在一起作为输入, 经过线性层和 *softmax* 函数, 最终输出两种嵌入对应的权重因子$\alpha$和$\beta$. 形式化地:

$$\alpha,\beta=softmax(linear(E_t \oplus E_m)) \tag{11}$$

其中, $\oplus$为向量拼接操作.

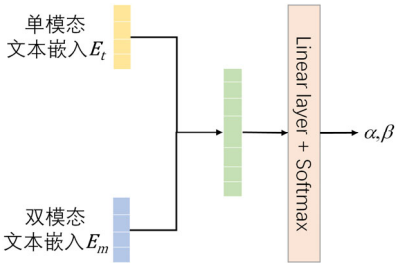



图 5 去噪网络模型架构

由于本文使用元学习方法学习通用的去噪网络, 因此, 本文将线性层的参数作为元参数$\Phi$, 根据公式(6), 使得模型通过大量源领域学习到最优参数$\Phi^*$. 接下来对去噪网络的元参数优化过程进行详细的介绍. 由于去噪网络用于发射模块的原型获取过程中, 因此, 本文对整个发射模块的元参数优化过程进行阐述. 图6展示了发射模块元参数整体优化过程.

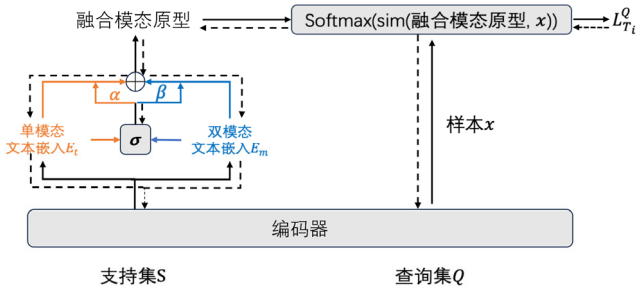



图 6 发射模块的元参数优化过程

在内层优化时, 如图 6 中的实线所示, 支持集样本首先经过编码器得到单模态文本嵌入和双模态文本嵌入. 然后将两种嵌入输入去噪网络, 得到对应的权重因子$\alpha$和$\beta$. 最后, 基于$\alpha$和$\beta$对单模态文本嵌入和双模态文本嵌入加权求和得到融合模态嵌入(公式(8)), 进而得到融合模态原型. 值得注意的是: 由于原型网络使用相似度计算的方式代替分类层计算发射分数, 所以在本文方法中, 内层不需要优化参数, 得到融合模态原型即可. 在外层优化时, 整个优化过程如图 6 中的虚线所示, 我们使用查询集评估模型性能, 即, 通过计算损失调整元参数并得到最优参数$\Phi^*$. 具体地, 基于内层优化得到的融合模态原型, 使用查询样本与融合原型计算相似度, 进而计算出损失$L_{T_i}^Q$. 然后反传$L_{T_i}^Q$来更新去噪网络和编码器的参数. 在所有源领域数据上重复上述操作, 最后得到最优元参数.

另外, 在整个训练过程中, 噪声数据并未区分定义, 我们将噪声数据与正常数据一同训练, 通过目标任务的学习, 让去噪网络判断每条样本中图像信息的有益程度, 从而减少图像信息中噪声数据的影响. 如果对噪声数据进行单独区分, 在训练过程中, 还需要让模型判断样本是否属于噪声数据, 这样会给模型增加额外的任务, 在小样本的情况下, 更加容易造成过拟合.

最后, 本文给出 MFNER 模型方法的伪代码, 如算法 1 所示. 首先, 从每个任务$T_i$中随机初始化出支持集$S_i$和查询集$Q_i$支持, 并将其输入编码器中得到文本嵌入和图像嵌入(第 4、5 行). 在内层优化时, 我们将两种嵌入拼接并输入去噪网络$g_{\Phi_n}$, 得到对应的权重因子, 根据权重因子加权求和得到融合嵌入, 进而得到融合模态$P_f$(第 6–8 行). 此外, 利用 CDT 机制将抽象转移矩阵$\Phi_m$展开成适应具体领域任务的转移矩阵$M_{expand}$(第 9 行). 在外层优化时, 给定所有查询集$Q_i$, 根据得到的$P_f^i$和$M_{expand}^i$计算元参数$\Phi$(第 11 行).

**算法 1**. MFNER 方法.

输入: 源领域$T$, 用于外层优化的学习率$\eta$.

输出: 元参数$\Phi$、编码器参数$\Phi_e$、去噪网络参数$\Phi_n$和抽象转移矩阵$\Phi_m$.

1. $\Phi$←随机初始化
2. **WHILE** 没有收敛
3. **FOR** 每个任务$T_i \in T$
4. $S_i, Q_i$←从$T_i$中随机采样
5. $E_t, E_m \leftarrow encoder_{\Phi_e}(S)$ //计算文本嵌入和多模态嵌入
6. $\alpha, \beta \leftarrow g_{\Phi_n}(E_t, E_m)$
7. $E_f \leftarrow \alpha \times E_t + \beta \times E_m$ //得到融合模态嵌入
8. $P_f^i \leftarrow avg(E_f)$ //得到融合模态原型

9. $M^{i}_{expand} \leftarrow CDT(\Phi_m)$ //得到适应任务的标签依赖矩阵

10. **END FOR**

11. $\Phi = \Phi - \eta \nabla_{\Phi} \sum_{T_i \in T} L^{Q_i}_{T_i}(P^{i}_{f}, M^{i}_{expand})$

12. **END WHILE**

13. 输出编码器参数$\Phi_e$、去噪网络参数$\Phi_n$和抽象转移矩阵$\Phi_m$

## 4 实验分析

### 4.1 实验数据

本文在 3 个公开的、由文本数据和图像数据构成的多模态数据集上进行实验. 3 个数据集分别如下.

(1) Twitter-15 数据集: 由 Zhang 等人[39]构建得到, 包括 4 个标签, 共有 8 257 个“文本+图像”样本对, 在构造成小样本数据集时, 1-shot 场景下有 200 个小样本, 5-shot 场景下有 172 个 few-shot 样本.

(2) SNAP 数据集: 由 Lu 等人[40]标注得到, 包括 4 个标签和 7 181 个“文本+图像”样本对, 在构造好的小样本数据集中, 1-shot 场景下有 200 个 few-shot 样本, 5-shot 场景下有 111 个 few-shot 样本.

(3) Twitter-17 数据集: SNAP 的过滤版, 由 Yu 等人[41]收集, 包括 4 个标签, 共有 4 819 个“文本+图像”样本对, 在构造的小样本数据集中, 1-shot 场景下有 200 个 few-shot 样本, 5-shot 场景下有 80 个 few-shot 样本.

为了模拟小样本场景, 本文根据 Hou 等人[23]提出的采样方法, 将上述 3 个原始数据集构建成 1-shot 和 5-shot 小样本数据集. 具体地, 我们根据“*N*-way *K*-shot”原则进行 few-shot 样本构建. 由于命名实体识别任务比较特殊, 即一条样本中可能出现多个实体, 所以我们规定 few-shot 样本中一个标签的样本个数不少于 *K* 个. 并在保证 few-shot 样本遵循“*N*-way *K*-shot”原则的前提下, 要求删除样本集合内的任意样本时, 该原则都不能被满足, 以保证样本是极少量的. 数据集的具体细节见表 2.

表 2 实验数据集

| 数据集 | 标签 | 样本数量 | *k*-shot | few-shot 样本数量 |
|---|---|---|---|---|
| Twitter-15 | ORG<br>OTHER<br>PER<br>LOC | 8 257 | 1-shot 或 5-shot | 200 或 172 |
| SNAP | ORG<br>MISC<br>PER<br>LOC | 7 181 | 1-shot 或 5-shot | 200 或 111 |
| Twitter-17 | ORG<br>MISC<br>PER<br>LOC | 4 819 | 1-shot 或 5-shot | 200 或 80 |

需要注意的是: 在构建 1-shot 小样本数据集时, 3 个数据集的 few-shot 样本数量均超过 200 个, 但由于受到设备和时间限制, 本文将小样本的数量控制在了 200 个以内.

### 4.2 评价指标及基准模型

本文采用评价指标 *F*1 值来评估小样本命名实体识别模型的性能. 作为常用的评价指标, *F*1 值被广泛地应用于之前的实体识别工作[42,43]中. *F*1 值是精确率(precision)和召回率(recall)的调和平均数. 针对某个类别的预测, 精确率是指被分类器判定为正例中的正样本与判定为正例的总样本的比例, 召回率指的是被预测为正例的正样本数占该类总的正例的比例. 具体公式如下.

$$F1 = 2 \times \frac{\text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \tag{12}$$

本文将所提模型 MFNER 与单模态小样本命名实体识别模型(包括 MNet[44]、WPZ[45]和 SpanNER[46])以及

多模态命名实体识别方法(ITA[22]和 Bert+CRF[47])进行了比较.

通过与单模态小样本命名实体识别模型进行比较, 可以证明多模态数据对小样本命名实体识别模型预测的有效性. 为此, 本文与基于原型网络的 MNet 和 WPZ 进行了比较, 它们都是用通过计算实体与原型簇相似度的方法进行实体预测. 其中: MNet 将每个类别的实体嵌入的平均作为对应类的原型构建原型网络; WPZ 使用了和 MNet 相同的策略, 但是将原型网络替换为匹配网络. 同时, 本文还与最新的、考虑了标签依赖的方法 SpanNER 进行了比较. SpanNER 将模型分为实体抽取与实体分类. 通过这种方式, 弱化 token-level 的标签依赖对模型的影响.

另外, 本文还与多模态实体识别方法 ITA 进行了比较. ITA 将图像信息转换为文本信息, 进而将图像信息有效映射在文本语义空间内, 实现多模态数据间的有效融合. 此外, 我们还与深度学习模型 BERT+CRF 进行了比较, BERT 是主流的基于神经网络的文本分类模型, 在大规模训练数据的支持下有强大的预测性能, CRF 框架主要用于学习实体识别任务中的标签依赖关系.

### 4.3 实验方法

参考之前的工作[23,43,44,46], 将 Twitter-15、SNAP 和 Twitter-17 这 3 个数据集构建为 1-shot 和 5-shot 的小样本数据集分别进行实验. 本文随机选取两个数据集作为训练集, 剩余的一个数据集作为测试集. 为了减少实验随机性造成的误差, 设置了 5 个不同的随机种子, 基于不同的随机种子进行实验, 并将 5 次实验结果的 $F1$ 值的平均值作为衡量指标.

本文基于操作系统 Ubuntu18.04, Python3.8 以及 PyTorch1.7, 并在 NVIDIA RTX 3090 GPU 进行实验. uncased BERT-Base[48]用于获取本文提出的方法以及基准方法的上下文嵌入. 此外, 实验使用 Adam[49]优化器训练模型. 其中, batchsize 为 4, 超参数学习率的选取范围是{1e–4,1e–5,1e–6}, CDT 中转移矩阵的更新学习率选取范围是{1e–2,1e–3,1e–4}, 去噪网络的学习率选取范围是{1e–3,1e–4,1e–5,1e–6,1e–7}. 对于每个模型的同一种参数配置, 采用的策略是以其能在所有的数据集取得最优平均值的参数组合作为最佳的参数配置.

### 4.4 实验结果与分析

为了评估基于多模态数据的小样本命名实体识别方法 MFNER 的有效性, 本文研究了以下两个问题.

- RQ1: MFNER 是否可以通过借助多模态数据获得比现有的先进小样本命名实体识别方法更好的结果?
- RQ2: 在训练数据十分受限的情况下, MFNER 是否比其他多模态命名实体识别方法更好?

为了评估验证上述两个问题的结果, 本文将 MFNER 模型分别与小样本实体识别方法和多模态实体识别方法在 1-shot 和 5-shot 两种小样本场景下进行了对比实验. 接下来针对每个问题进行实验结果分析.

#### 4.4.1 RQ1: MFNER 是否可以通过借助多模态数据获得比现有的先进小样本命名实体识别方法更好的结果?

为了验证这个问题的结果, 本文将 MFNER 模型与现有的先进小样本实体识别方法 MNet, WPZ 和 SpanNER 进行比较. 其中, MNet 和 WPZ 未考虑标签之间的依赖关系, SpanNER 弱化了标签依赖对预测结果的影响. 实验的结果见表 3.

与不考虑标签依赖的原型方法 MNet 和 WPZ 相比, 本文提出的方法 MFNER 有着巨大的优势, 在 1-shot 和 5-shot 的情况下, 在所有的数据集上都获得了很好的效果, 尤其在 1-shot 的小样本情况下, 平均 $F1$ 值上至少提高了 14%. 对于这一结果推测如下: 在数据量极少的情况下, 首先, 传统原型网络中不同领域的原型会互相干扰造成模型误分类, 在这种情况下, 基于学好的原型网络, 在预测时将不同领域的原型映射到不同的空间中, 可以有效避免干扰, 提高模型的预测效果; 其次, 由于训练数据十分有限, 其包含的语义信息可能也是受限的, 进而导致无法学习有效的语义空间. 为此, 加入多模态数据信息来提供额外的有效语义信息, 可以帮助模型学习更加丰富的语义空间, 提高小样本模型的泛化性能. 在 5-shot 的情况下, MNet 的 $F1$ 值略高于 MFNER. 导致这个结果的原因可能是: Twitter-17 是 SNAP 的过滤版, 因此两个数据集含有丰富的共同语义知识, 在对 SNAP 进行测试时, Twitter-17 作为训练集, 模型可以从 Twitter-17 中迁移大量共性知识用于模型在

SNAP 数据集上进行学习. 但是 MFNER 中使用的 L-TapNet 将不同的领域映射在不同的语义空间, 阻碍了这种共性知识的学习; 而 MNet 构建的语义空间中, 所有领域共享同一语义空间, 可以更好地利用上述共性知识.

表 3 MFNER 与现有方法在 3 个数据集上的 *F*1 值(给出了标准差)

| *k*-shot | 方法 | Twitter-17 | Twitter-15 | SNAP | Avg. |
|---|---|---|---|---|---|
| 1-shot | MNet | 77.47±0.50 | 55.22±0.78 | 73.83±0.26 | 68.84±0.38 |
| | WPZ | 67.13±0.98 | 53.36±0.49 | 65.73±0.73 | 62.07±0.29 |
| | SpanNER | 75.54±0.33 | 47.02±0.22 | 73.41±031 | 65.32±0.20 |
| | BERT+CRF | 0.88±0.87 | 0.78±0.69 | 1.02±1.04 | 1.62±2.12 |
| | ITA | 0.88±0.88 | 0.80±0.70 | 1.01±1.04 | 1.67±2.10 |
| | MFNER (ours) | **82.58±0.43** | **66.79±0.33** | **80.04±0.57** | **76.47±0.35** |
| 5-shot | MNet | **84.36±0.29** | 67.67±0.27 | **76.87±0.20** | **76.30±0.19** |
| | WPZ | 76.61±0.85 | 66.22±0.70 | 71.99±0.84 | 71.61±0.38 |
| | SpanNER | 74.77±0.26 | 50.67±0.32 | 67.97±0.22 | 64.47±0.17 |
| | BERT+CRF | 1.09±1.06 | 0.95±0.75 | 1.30±1.27 | 1.11±0.92 |
| | ITA | 0.94±0.96 | 0.95±0.82 | 1.10±1.04 | 1.00±0.75 |
| | MFNER (ours) | 82.93±0.64 | **68.60±0.21** | 73.87+0.67 | 75.13+0.38 |

与考虑但弱化了标签依赖对模型影响的原型方法 SpanNER 相比, 从 3 个数据集的预测结果上看, 无论是 1-shot 还是 5-shot 的情况下, MFNER 均超越了 SpanNER. 其中: 在 1-shot 情况下的数据集 Twitter-15 上, MFNER 的 *F*1 值提高了 19%. 为此, 本文做出如下分析: SpanNER 通过“BIO”标签将候选实体提取出来, 进而避免考虑 token-level 的标签依赖转移. SpanNER 忽略了标签的语义信息, 但是对于学习语义空间来讲, 标签信息也是十分重要的. 因此, 考虑了标签信息的 token-level 标签依赖转移, 可以更加有效地帮助小样本模型提高泛化性.

综上, MFNER 显著优于现有的小样本命名实体识别方法且更具有实用价值.

4.4.2 RQ2: 在训练数据十分受限的情况下, MFNER 是否比其他多模态命名实体识别方法更好?

为了回答该问题, 本文将所提方法与现有的多模态方法和深度学习方法的性能进行了比较, 包括 BERT 和 ITA. 值得注意的是: BERT 是通用的深度学习模型, 本文将图像转成文本后与原始文本拼接作为样本, 输入 BERT 进行实体识别. BERT 和 ITA 都是在与本文相同实现配置下进行的比较, 实验结果见表 3. 从结果可以看出, MFNER 的性能令人印象深刻. 在小样本数据集的情况下, ITA 和 BERT+CRF 模型的表现都十分不佳. 可能的原因为: ITA 和 BERT+CRF 模型都是用 BERT 模型作为嵌入模型, 需要大量数据预训练以保证预测效果. 而在小样本场景下, 任务 $T_i$ 的样本集为 few-shot 样本, 因此, 在预测前使用小样本的支持集(只有十几条样本)对模型进行微调, 导致模型严重过拟合, 使得 ITA 和 BERT+CRF 模型的泛化性能十分受限.

因此, 综合而言, 在数据量十分受限的情况下, MFNER 相对于现有的多模态命名实体识别方法有较大优势, 并更具有指导意义和实践价值.

### 4.5 消融实验

为了充分说明考虑多模态数据的有效性以及去噪网络的必要性, 本文在 1-shot 的小样本情况下分别设置了两种场景用于分析上述两种情况对性能带来的影响.

首先, 为了证明考虑图像作为辅助模态的确可以帮助小样本模型提高泛化性能, 本文将不考虑图像的单模态文本数据作为输入, 使 MFNER 在没有辅助模态帮助的情况下, 进行命名实体识别任务. 表 4 分别展示了使用图像数据和不使用图像数据的结果. 在 3 个数据集上, 使用图像数据辅助模型预测时, 每个数据的 *F*1 值都有较大的提升. 这说明考虑多模态数据的确可以帮助模型构建更丰富的语义空间, 进而提高小样本模型的泛化性能.

为了进一步说明辅助模态数据的有效性, 本文给出了具体实例分析. 如图 7 所示, 对于样本中的最后一个词“Katie”, 它的真实标签为“B-PER(person 的缩写)”. 在使用单模态数据的情况下, “Katie”被错误识别为“O”; 在使用多模态数据的情况下则被正确识别. 本文分析原因如下: 对于只使用单模态数据的情况下, 由于

“Katie”前一个词为“so”，结合上下文语义信息，“Katie”很容易被识别为形容词而不是名词“person”. 而在加入了图像文本的情况下，由于图像文本中频繁出现表示“person”的信息，例如“woman”和“her”，为模型额外提供了正确的语义信息，进而帮助模型做出了正确的预测. 这进一步说明辅助模态信息的确可以帮助提升小样本模型的预测性能.

表 4 多模态数据相关的消融实验

| *k*-shot | 方法 | Twitter-17 | Twitter-15 | SNAP | Avg. |
|---|---|---|---|---|---|
| 1-shot | 使用单模态数据 | 80.78±0.24 | 64.73±0.78 | 79.01±0.66 | 75.47±0.42 |
| | 使用多模态数据 | **82.33±0.47** | **66.18±0.60** | **79.61±0.90** | **76.04±0.57** |

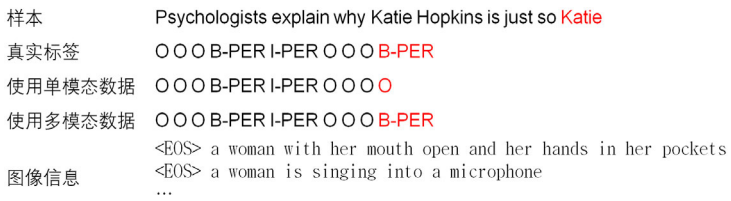



图 7 多模态融合实验具体实例

其次，为了证明 MFNER 中去噪网络的必要性，本文设置了使用去噪网络和不使用去噪网络两种实验进行对比，其实验结果见表 5. 当模型使用去噪网络时，在 3 个数据集上的预测 *F*1 值均有提升. 这个结果证明了去噪网络的有效性，说明在借助其他模态数据时，样本中的确存在噪音数据，而使用去噪网络可以有效地帮助模型评估辅助模态数据的有益程度.

表 5 去噪网络相关的消融实验

| *k*-shot | 方法 | Twitter-17 | Twitter-15 | SNAP | Avg. |
|---|---|---|---|---|---|
| 1-shot | 未使用去噪网络 | 82.33±0.47 | 66.18±0.60 | 79.61±0.90 | 76.04±0.57 |
| | 使用去噪网络 | **82.58±0.43** | **66.79±0.33** | **80.04±0.57** | **76.47±0.35** |

为了进一步证明去噪网络的必要性，本文给出了具体实例分析. 如图 8 所示：对于样本中的词“Isaiah”，它的真实标签为“B-PER”. 在未使用去噪网络的情况下，“Isaiah”被错误识别为“O”；在使用去噪网络的情况下则被正确识别. 本文分析原因如下：由于去除了标点符号，“Isaiah”的上下文信息并不是很明确，实际上，“Isaiah”是以主人公名字命名的一本书. 因此，“Isaiah”作为实体，既可以表示“person”也可以表示“book”. 由真实标签可以看出，此处“Isaiah”更偏向于被识别为“person”. 然而在加入图像信息后，由于图像信息中提供的语义都与“book”更相关，例如“writing”和“font”，所以模型更偏向于认为“Isaiah”不是“person”，进而导致模型预测错误. 在这种情况下，图像信息作为噪声样本对模型产生了错误的引导，而通过使用去噪网络缓解噪声样本对模型的影响后，模型可以正确识别出“Isaiah”的标签. 这进一步说明：去噪网络的确可以帮助模型有效评估不同样本的有益程度，并缓解噪声样本对模型的不良影响.

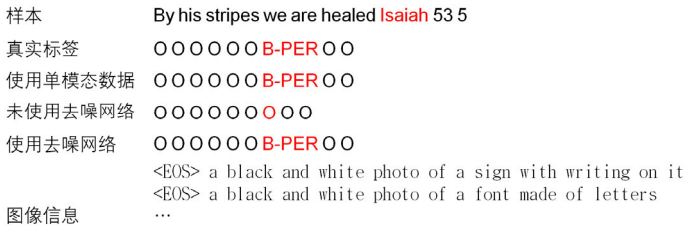



图 8 去噪网络消融实验具体实例

总体而言，借助图像数据作为辅助模态，的确可以帮助面向文本的小样本命名实体识别模型提高泛化性能. 此外，对辅助模态进行去噪操作是十分必要的，去噪网络可以有效筛选出对模型有益的样本，实现多模态

数据间的有效融合, 进而达到辅助模型的目的.

### 4.6 分析实验

为了进一步说明多模态数据可以帮助小样本场景的命名实体识别模型提高泛化性能, 本文设计了如下实验: 将只考虑单模态文本信息的小样本 SOTA 方法和引入视觉辅助信息的小样本 SOTA 方法进行对比. 其中, 对于引入视觉辅助信息的小样本 SOTA 方法, 本文将文本和图像转文本拼接作为模型输入, 然后利用 BERT 模型获取考虑了图像转文本信息的文本嵌入, 并将其作为多模态嵌入用于训练和测试.

表 6 展示了在 1-shot 情况下的实验结果. 从实验结果可以看出: 在加入多模态信息后, MNet 和 WPZ 模型的预测效果均有提升, 其中, WPZ 模型提升非常明显, 在 3 个数据集上, *F*1 值分别至少提升了 11%、3%和 9%. 说明图像信息的确起到了为小样本数据提供额外语义的作用, 进而证明了多模态数据对小样本场景的有效性. SpanNER 模型在加入视觉辅助信息后, 在 3 个数据集上的预测效果均有下降. 这可能的原因是: 在 SpanNER 的实体抽取阶段, 候选实体只需要被预测为标签 B, I 和 O, 并未考虑到包含具体语义信息的标签信息; 而加入视觉辅助信息作为额外语义信息后, 没有标签信息的引导, 反而对实体抽取造成了干扰. MFNER 和加入了多模态信息的 SOTA 方法相比, 依然在 3 个数据集上均表现最优, 说明了 MFNER 强大的泛化性能.

表 6 多模态与单模态的分析实验

| *k*-shot | 方法 | | Twitter-17 | Twitter-15 | SNAP | Avg. |
|---|---|---|---|---|---|---|
| 1-shot | MNet | 单模态 | 77.47±0.50 | 55.22±0.78 | 73.83±0.26 | 68.84±0.38 |
| | | 多模态 | **77.56±0.27** | **56.08±0.48** | **73.94±0.51** | **69.19±0.19** |
| | WPZ | 单模态 | 67.13±0.98 | 53.36±0.49 | 65.73±0.73 | 62.07±0.29 |
| | | 多模态 | **78.77±0.33** | **56.61±0.50** | **74.92±0.44** | **70.10±0.16** |
| | SpanNER | 单模态 | **75.54±0.33** | **47.02±0.22** | **73.41±031** | **65.32±0.20** |
| | | 多模态 | 71.30±1.05 | 41.98±0.64 | 69.92±0.36 | 61.06±0.56 |
| | MFNER (ours) | | **82.58±0.43** | **66.79±0.33** | **80.04±0.57** | **76.47±0.35** |

## 5 总 结

本文提出了一种融合多模态数据的小样本命名实体识别模型 MFNER, 针对面向文本的小样本命名实体识别任务, 利用图像模态数据作为辅助模型帮助模型提升泛化性能. MFNER 模型使用 CRF 框架并主要分为 3 个模块: 多模态信息提取模块、发射模块和转移模块. 多模态信息提取模块对图像数据进行预处理, 将其转化为文本信息, 以实现不同模态嵌入的有效对齐, 从而达到辅助的目的; 发射模块使用先进的基于原型网络的方法学习通用的语义空间, 将查询样本与原型的相似度作为发射分数; 转移模块通过学习通用的抽象标签依赖转移, 使模型有效地捕捉标签之间的依赖关系. 此外, 本文提出了基于元学习的去噪网络, 通过去噪网络, 缓解图像数据中噪声样本对模型预测产生错误引导的现象. 最后, 在真实的单模态和多模态数据集上进行了全面的实验, 实验结果验证了本文所提出的 MFNER 模型的有效性.

**References**:

[1] Mikheev A, Moens M, Grover C. Named entity recognition without gazetteers. In: Proc. of the 9th Conf. of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1999. 1–8.

[2] Fei H, Wu SQ, Li JY, *et al*. LasUIE: Unifying information extraction with latent adaptive structure-aware generative language model. In: Proc. of the 26th Conf. on Neural Information Processing Systems. 2022. 15460–15475.

[3] Schäfer H, Idrissi-Yaghir A, Horn P, *et al*. Cross-language transfer of high-quality annotations: Combining neural machine translation with cross-linguistic span alignment to apply NER to clinical texts in a low-resource language. In: Proc. of the 4th Clinical Natural Language Processing Workshop. Stroudsburg: Association for Computational Linguistics, 2022. 53–62.

[4] Bień M, Gilski M, Maciejewska M, *et al*. RecipeNLG: A cooking recipes dataset for semi-structured text generation. In: Proc. of the 13th Int'l Conf. on Natural Language Generation. Stroudsburg: Association for Computational Linguistics, 2020. 22–28.

[5] Fan RY, Wang LZ, Yan JN, *et al*. Deep learning-based named entity recognition and knowledge graph construction for geological hazards. ISPRS Int'l Journal of Geo-information, 2019, 9(1): 15−37.

[6] Wu YZ, Li HR, Yao T, *et al*. A survey of multimodal information processing frontiers: Application, fusion and pre-training. Journal of Chinese Information Processing, 2022, 36(5): 1−20 (in Chinese with English abstract).

[7] Sun L, Wang JQ, Zhang K, *et al*. RpBERT: A text-image relation propagation-based BERT model for multimodal NER. In: Proc. of the AAAI Conf. on Artificial Intelligence. Palo Alto: Association for the Advancement of Artificial Intelligence, 2021. 13860−13868. [doi: 10.3969/j.issn.1003-0077.2022.05.001]

[8] Liu W, Xu TG, Xu QH, *et al*. An encoding strategy based word-character LSTM for Chinese NER. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019. 2379−2389. [doi: 10.18653/v1/N19-1247]

[9] Jia C, Zhang Y. Multi-cell compositional LSTM for NER domain adaptation. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020. 5906−5917.

[10] Chang Y, Kong L, Jia K, *et al*. Chinese named entity recognition method based on BERT. In: Proc. of the 2021 IEEE Int'l Conf. on Data Science and Computer Application (ICDSCA). Piscataway: IEEE, 2021. 294−299.

[11] Ju SG, Li TN, Sun JP. Chinese fine-grained name entity recognition based on associated memory networks. Ruan Jian Xue Bao/ Journal of Software, 2021, 32(8): 2545−2556 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6114.htm [doi: 10.13328/j.cnki.jos.006114]

[12] Zhao SJ. Named entity recognition in biomedical texts using an HMM model. In: Proc. of the Int'l Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. Stroudsburg: Association for Computational Linguistics, 2004. 84−87.

[13] Zhou GD, Su J. Named entity recognition using an HMM-based chunk tagger. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002. 473−480.

[14] Konkol M, Konopík M. CRF-based Czech named entity recognizer and consolidation of Czech NER research. In: Proc. of the 16th Int'l Conf. on Text, Speech, and Dialogue. Berlin: Springer, 2013. 153−160.

[15] Liu J, Chen Y, Xu J. Low-resource NER by data augmentation with prompting. In: Proc. of the 31st Int'l Joint Conf. on Artificial Intelligence. 2022. 4252−4258.

[16] Ma RT, Zhou X, Gui T, *et al*. Template-free prompt tuning for few-shot NER. In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2022. 5721−5732. [doi: 10.18653/v1/2022.naacl-main.420]

[17] Mettes P, van der Pol E, Snoek CGM. Hyperspherical prototype networks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. 2019. 1487−1497.

[18] Li G, Jampani V, Sevilla-Lara L, *et al*. Adaptive prototype learning and allocation for few-shot segmentation. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021. 8334−8343.

[19] Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 327−348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[20] Yin J, Zhang ZD, Gao YH, Yang ZW, Li L, Xiao M, Sun YQ, Yan CG. Survey on vision-language pre-training. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2000−2023 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6774.htm [doi: 10.13328/j.cnki.jos.006774]

[21] Chen SG, Aguilar G, Neves L, *et al*. Can images help recognize entities? A study of the role of images for Multimodal NER. In: Proc. of the 7th Workshop on Noisy User-generated Text (W-NUT 2021). Stroudsburg: Association for Computational Linguistics, 2021. 87−96. [doi: 10.18653/v1/2021.wnut-1.11]

[22] Wang XY, Gui M, Jiang Y, *et al*. ITA: Image-text alignments for multi-modal named entity recognition. In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2022. 3176−3189.

[23] Hou YT, Che WX, Lai YK, *et al*. Few-shot slot tagging with collapsed dependency transfer and label- enhanced task-adaptive projection network. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020. 1381−1393. [doi: 10.18653/v1/2020.acl-main.128]

[24] Sui DB, Tian ZK, Chen YB, *et al*. A large-scale Chinese multimodal NER dataset with speech clues. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021. 2807−2818. [doi: 10.18653/v1/2021.acl-long.218]

[25] Meng YX, Wu W, Wang F, *et al*. Glyce: Glyph-vectors for Chinese character representations. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. 2019. 2746−2757.

[26] Wu S, Song XN, Feng ZH. MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021. 1529−1539. [doi: 10.18653/v1/2021.acl-long.121]

[27] Moon S, Neves L, Carvalho V. Multimodal named entity recognition for short social media posts. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018. 852−860. [doi: 10.18653/v1/N18-1078]

[28] Sun L, Wang JQ, Su YD, *et al*. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020. 1852−1862.

[29] Pahde F, Puscas M, Klein T, *et al*. Multimodal prototypical networks for few-shot learning. In: Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision. Piscataway: IEEE, 2021. 2644−2653. [doi: 10.1109/WACV48630.2021.00269]

[30] Memmesheimer R, Theisen N, Paulus D. SL-DML: Signal level deep metric learning for multimodal one-shot action recognition. In: Proc. of the 25th Int'l Conf. on Pattern Recognition (ICPR). Piscataway: IEEE, 2021. 4573−4580.

[31] Aktukmak M, Yilmaz Y, Hero A. Any-shot learning from multimodal observations (ALMO). IEEE Access, 2023, 11: 61513−61524.

[32] Lin ZQ, Yu S, Kuang ZY, *et al*. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023. 19325−19337.

[33] Tsimpoukelli M, Menick JL, Cabi S, *et al*. Multimodal few-shot learning with frozen language models. In: Proc. of the 35th Conf. on Neural Information Processing Systems. 2021. 200−212.

[34] Wan H, Zhang M, Du JF, *et al*. FL-MSRE: A few-shot learning based approach to multimodal social relation extraction. In: Proc. of the AAAI Conf. on Artificial intelligence. Palo Alto: Association for the Advancement of Artificial Intelligence, 2021. 13916−13923. [doi: 10.1609/aaai.v35i15.17639]

[35] Alayrac JB, Donahue J, Luc P, *et al*. Flamingo: A visual language model for few-shot learning. In: Proc. of the 36th Conf. on Neural Information Processing Systems. 2022. 23716−23736.

[36] Ma X, Hovy E. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016. 268−278.

[37] Forney GD. The Viterbi algorithm. Proc. of the IEEE, 61(3): 268−278. [doi: 10.1109/PROC.1973.9030]

[38] Ding N, Xu GW, Chen YL, *et al*. Few-NERD: A few-shot named entity recognition dataset. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing, Vol. 1 (Long Papers). Stroudsburg: Association for Computational Linguistics, 2021. 3198−3213.

[39] Zhang Q, Fu JL, Liu XY, *et al*. Adaptive co-attention network for named entity recognition in Tweets. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. Palo Alto: Association for the Advancement of Artificial Intelligence, 2018. 5674−5681. [doi: 10.1609/aaai.v32i1.11962]

[40] Lu D, Neves L, Carvalho V, *et al*. Visual attention model for name tagging in multimodal social media. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018. 1990−1999.

[41] Yu JF, Jiang J, Yang Li, *et al*. Improving multimodal named entity recognition via entity Span detection with unified multimodal transformer. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020. 3342−3352. [doi: 10.18653/v1/2020.acl-main.306]

[42] Fang JY, Wang XB, Meng ZQ, *et al*. MANNER: A variational memory-augmented model for cross domain few-shot named entity recognition. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics Vol. 1 (Long Papers). Stroudsburg: Association for Computational Linguistics, 2023. 4261−4276.

[43] Chen JW, Liu Q, Lin HY, *et al*. Few-shot named entity recognition with self-describing networks. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022. 5711−5722.

[44] Vinyals O, Blundell C, Lillicrap T, *et al*. Matching networks for one shot learning. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. 2016. 3637−3645.

[45] Fritzler A, Logacheva V, Kretov M. Few-shot classification in named entity recognition task. In: Proc. of the 34th ACM/SIGAPP Symp. on Applied Computing. New York: Association for Computing Machinery, 2019. 993−1000.

[46] Ma TT, Jiang HQ, Wu QH, *et al*. Decomposed meta-learning for few-shot named entity recognition. In: Proc. of the Findings of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022. 1584−1596.

[47] Souza F, Nogueira R, Lotufo R. Portuguese named entity recognition using BERT-CRF. arXiv:1909.10649, 2019.

[48] Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the NAACL-HLT. Stroudsburg: Association for Computational Linguistics, 2019. 4171−4186. [doi: 10.18653/v1/N19-1423]

[49] Diederik PK, Jimmy B. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

**附中文参考文献**:

[6] 吴友政, 李浩然, 姚霆, 等. 多模态信息处理前沿综述: 应用、融合和预训练. 中文信息学报, 2022, 36(5): 1−20.

[11] 琚生根, 李天宁, 孙界平. 基于关联记忆网络的中文细粒度命名实体识别. 软件学报, 2021, 32(8): 2545−2556. http://www.jos.org.cn/1000-9825/6114.htm [doi: 10.13328/j.cnki.jos.006114]

[19] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述. 软件学报, 2021, 32(2): 327−348. http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[20] 殷炯, 张哲东, 高宇涵, 杨智文, 李亮, 肖芒, 孙垚棋, 颜成钢. 视觉语言预训练综述. 软件学报, 2023, 34(5): 2000−2023. http://www.jos.org.cn/1000-9825/6774.htm [doi: 10.13328/j.cnki.jos.006774]

**张天明**(1988－), 女, 博士, 讲师, CCF 专业会员, 主要研究领域为图数据管理, 深度学习.

**张杉**(1995－), 女, 博士生, CCF 学生会员, 主要研究领域为机器学习, 自然语言处理, 小样本学习.

**刘曦**(1999－), 男, 硕士生, CCF 学生会员, 主要研究领域为机器学习, 自然语言处理.

**曹斌**(1985－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为时空数据库, 数据挖掘.

**范菁**(1969－), 女, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为中间件, 虚拟现实, 可视化.