CLGLF:置信学习引导标签融合的多模态命名实体识别方法

王海荣[1,2],王 彤[1*],徐 玺[1],荆博祥[1],陈芳萍[1]
(1. 北方民族大学计算机科学与工程学院,宁夏银川 750021;
2. 北方民族大学图像图形智能处理国家民委重点实验室,宁夏银川 750021)

摘 要: 为解决多模态命名实体识别中存在的视觉语义理解和多模态语义的偏差问题,本文提出了置信学习引导标签融合的多模态命名实体识别方法. 该方法调用BLIP-2预训练模型生成图像描述,将其与输入的文本拼接,进行图文联合编码实现多模态特征融合,对多模态表征和文本表征解码后得到候选标签和文本标签;在采用KL散度损失函数对齐两组标签的基础上,计算置信分数用来评估多模态表征质量,设置置信阈值辅助筛选出有偏差的候选标签,并使用相应位置的文本标签替换有偏差的候选标签,实现标签的融合,最终完成多模态命名实体识别. 为了验证本文方法,在Twitter-2015和Twitter-2017多模态数据集上进行实验,并将实验结果与MSB、UMT等7种主流方法进行对比,实验结果证明了本文方法的有效性.

关键词: 多模态命名实体识别;图像描述;置信学习;多模态语义偏差;信息抽取
基金项目: 宁夏自然科学基金(No.2023AAC03316);北方民族大学研究生创新项目(No.YCX23159)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)07-2429-09
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20231160

CLGLF: Confidence Learning Guides Label Fusion for Multimodal Named Entity Recognition Method

WANG Hai-rong[1,2], WANG Tong[1*], XU Xi[1], JING Bo-xiang[1], CHEN Fang-ping[1]
(1. *School of Computer Science and Engineering, North Minzu University, Yinchuan, Ningxia* 750021, *China*;
2. *Laboratory of Image & Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan, Ningxia* 750021, *China*)

Abstract: To solve the visual semantic understanding bias and multimodal semantic bias in multimodal named entity recognition, the confidence learning guides label fusion (CLGLF) method for multimodal named entity recognition is proposed. This method invokes the BLIP-2 pre-trained model to generate image captions, concatenates them with the input texts, and performs joint coding to achieve multimodal feature fusion. The candidate labels and text labels are obtained after decoding the multimodal representations and text representations. Based on using the KL divergence loss function to align the two groups of labels, the confidence score is calculated to evaluate the quality of the multimodal representation, and a confidence threshold is set to help screen out the biased candidate labels, the text labels in the corresponding positions are used to replace the biased candidate labels, to achieve the label fusion, and finally complete the multimodal named entity recognition. In order to verify the proposed method, experiments are carried out on the Twitter-2015 and Twitter-2017 multimodal datasets, and the experimental results are compared with 7 mainstream methods, such as MSB and UMT. The experimental results show the effectiveness of the CLGLF.

Key words: multimodal named entity recognition; image caption; confidence learning; multimodal semantic bias; information extraction

Foundation Item(s): Natural Science Foundation of Ningxia Province (No.2023AAC03316); Graduate Innovation Program of North Minzu University (No.YCX23159)

收稿日期:2023-12-15;修回日期:2024-04-01;责任编辑:王天慧
*通讯作者:王彤

1 引言

针对命名实体识别任务中存在的文本语义不充分问题,以往的方法主要通过字符特征、知识图谱和语料库等辅助信息进行文本语义补充. 近年来,随着多媒体数据的爆炸式增长,学者们开始探索将图像和声学等多模态数据作为辅助信息,用来解决命名实体识别任务中的歧义性问题,致使多模态命名实体识别(Multimodal Named Entity Recognition,MNER)方法研究备受关注.

MNER的本质是如何学习有效的视觉特征,并将其整合到文本表征中. 丰富的视觉上下文信息,可以为文本提供更准确、更全面的线索,但引入的图像数据并不总是有益的,如图像与文本内容不匹配将无法实现语义关联,且图像中与文本不相关或弱相关的视觉内容作为噪声信息,可能导致实体类型的错误预测,故最大程度地挖掘图像语义信息、优化视觉表征是重要的. 进一步通过融合视觉特征和文本特征,可以弥补单一模态数据的局限性,从而提高MNER的精确度. 因此,MNER的研究大多聚焦于图像的语义信息挖掘和多模态特征融合[1]两个方面.

现有的MNER研究工作对视觉特征的提取进行了由粗粒度到细粒度的改进,以帮助提高标签预测的准确性. 早期方法,如MA[2]、VAM[3]、CWI[4]使用卷积神经网络对整个图像进行编码,但没有充分利用视觉特征;ACN[5]、CAT[6]、MLMNER[7]、RpBERT[8]等方法则采用ResNet提取区域视觉特征,获得了更丰富的视觉语义表示. 上述MNER方法虽然提取了图像的粗粒度特征,但没有考虑视觉对象和文本实体之间的映射关系,因此无法确保视觉语义的完整性. 为此,GAN[9]通过预训练模型,识别图像中的对象;UMGF[10]引入四个实体类别词辅助挖掘更多的视觉对象. 这些方法关注了图像的细粒度特征,改善了多模态信息的噪声问题,但仍无法解决图文特征之间存在的模态差异问题. 针对此问题,OCSGA[11]利用对象级视觉标签来消除模态间的语义差异,ITJMNER[12]、ITA[13]引入图像标题生成技术,将视觉特征映射到文本特征空间,进而缩小模态间的差异,但通常存在生成的视觉标签和图像描述与图像内容不一致的问题.

多模态数据中存在的图文特征级语义冲突和噪声问题,将引起多模态表征发生偏差或错误,本文将其统称为多模态语义偏差. 针对多模态语义偏差,学者们试图通过降低模态融合中的多模态数据噪声,达到提高多模态表征质量的目的,如HvpNet[14]将层次化视觉前缀作为文本提示,MAF[15]、TGF[16]、DGCF[17]和RGCN[18]则利用跨模态匹配减少无关图像信息对模型的影响;MNER-QG[19]调用检索方法获取实体类型和视觉区域的先验知识,从而实现文本实体与视觉区域之间的对齐;DebiasCL[20]采用硬样本挖掘策略和去偏对比学习,减少视觉对象与文本实体在数量和类型上的偏差. 这类方法虽然在一定程度上提高了多模态表征的质量,但增加了模型的复杂度,而且忽略了原始文本中的关键信息. 为此,一些研究在解码阶段将NER作为辅助任务来减少多模态语义偏差,提高标签序列的准确度. 例如,UMT[21]利用基于纯文本的实体跨度检测作为辅助模块指导最终的预测;UAMNer[22]从文本中生成候选标签,调用多模态Transformer改善不确定性较高的候选标签;ITA[13]利用跨模态输入视图,提高文本输入视图的准确性;M3S[23]通过融合命名实体分类和命名实体分割任务,减少视觉偏差.

上述方法从图像语义信息挖掘和多模态特征融合两个角度提升了MNER方法性能,但仍未解决视觉语义理解和多模态语义的偏差问题. 基于上述发现,本文提出了一种置信学习引导标签融合(Confidence Learning Guides Label Fusion,CLGLF)的多模态命名实体识别方法. 具体来说,使用BLIP-2[24]将视觉识别推进至视觉理解[25],生成高质量的图像描述,在获得多模态表征和文本表征后,调用CRF解码生成候选标签和文本标签,通过构造置信阈值引导的标签融合机制提升方法性能.

本文的主要贡献总结如下:

(1)提出一个新的MNER方法,该方法通过BLIP-2将视觉识别推进至视觉理解,在缩小图文间模态差异的同时,更精确地捕获了图像中的关键信息和复杂的图像语义,从而增强了视觉语义理解;

(2)构造了一个置信阈值引导的标签融合机制,该机制通过NER生成的文本标签辅助校正有偏差的候选标签,实现两组标签融合,有效地减少了多模态语义偏差对方法的消极影响;

(3)在Twitter-2015和Twitter-2017两个数据集上进行大量实验,结果表明所提出的CLGLF方法是一种有效且有竞争力的MNER方法.

2 CLGLF方法模型

对于给定的文本-图像对,构建一个端到端的MNER框架来检测目标模态中的命名实体,与以往方法不同的是,CLGLF聚焦于多模态语义偏差问题,利用当前较先进的视觉语言预训练模型BLIP-2优化视觉语义. 具体的,CLGLF方法调用BLIP-2预训练模型生成图像描述语句,并将得到的视觉嵌入与文本嵌入拼接作为新的跨模态输入表示,通过图文联合编码获取多模态表征. 在训练的过程中,使用KL散度损失函数对齐多模态表征和文本表征的预测向量,将两个预测向量分别作为CRF的输入,生成候选标签和文本标签. 最

重要的是,通过置信阈值引导的标签融合机制,融合候选标签和文本标签,最终生成结果标签,有效解决了KL散度损失函数对齐不完全的问题,同时降低了多模态语义偏差对MNER的影响. 方法框架如图1所示.

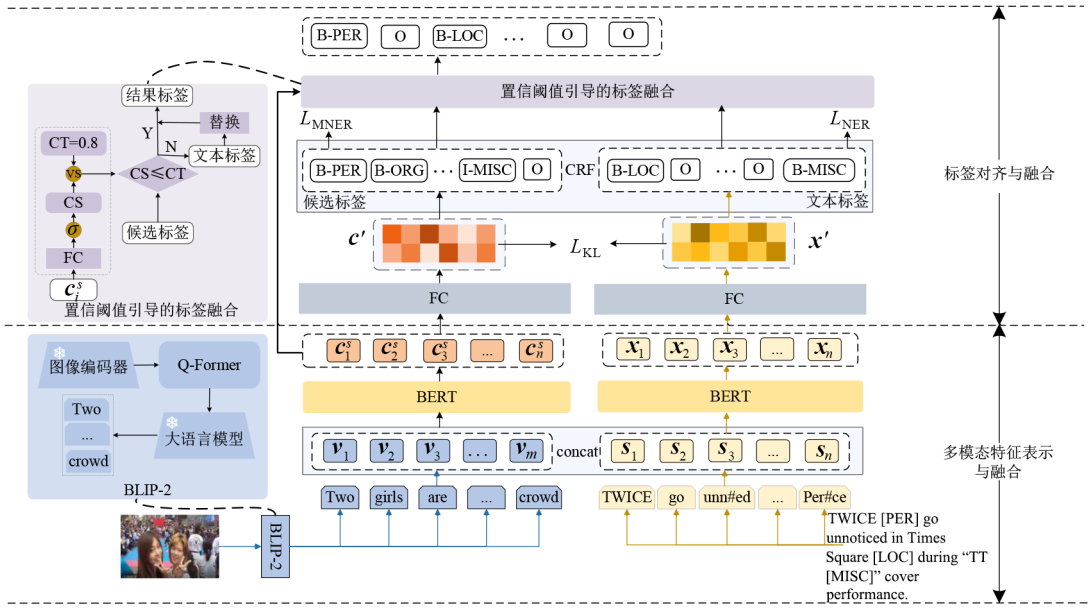


图1　CLGLF方法模型

图1展示了CLGLF方法的整体框架. 该方法主要包括多模态特征表示与融合、标签对齐与融合两个核心处理. 多模态特征表示与融合部分通过BLIP-2生成图像描述语句,将其与文本拼接后输入BERT进行图文联合编码,实现多模态特征融合;标签对齐与融合部分使用KL散度损失函数最小化多模态特征和文本特征的差距,并构造了一个置信阈值引导的标签融合策略,以减少MNER中的多模态语义偏差.

3　多模态特征表示与融合

3.1　多模态特征表示

针对文本数据,使用BERT预训练模型提取文本特征. 具体来说,调用预先训练好的BERT模型,把长度为n的文本嵌入$\boldsymbol{S}=[\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n]$编码为$d$维的实值向量,得到的文本表征可表示为$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$,其中,$\boldsymbol{X}\in\mathbb{R}^{n\times d}$.

针对图像数据,利用BLIP-2强大的视觉理解能力提取图像特征. BLIP-2是一个视觉语言预训练模型,它通过轻量级的Q-Former弥补了视觉和语言间的模态差异,实现了高效的图文语义对齐. 具体的,将图像输入到冻结的图像编码器中,强制Q-Former学习与文本最相关的视觉特征,接着将Q-Former的输出连接到冻结的大型语言模型中,执行由视觉到文本的生成学习. 与提取全局和区域视觉特征的视觉识别技术相比,BLIP-2生成的图像描述能够综合图像中的关键信息和复杂的图像语义,同时缩小了模态差距,利于特征融合. 此外,与以往的视觉标签和图像标题生成技术相比较,BLIP-2能够更精确地捕获图像中对象之间的所属关系、位置关系等,从而产生更准确的图像描述语句. 具体到本文方法,将图像$\boldsymbol{I}$输入到图像编码器中,通过Q-Former学习与文本最相关的视觉特征,将Q-Former的输出连接到大语言模型OPT[26]中,生成一条由m个单词组成的图像描述语句,进一步得到视觉嵌入$\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m]$,其中,$\boldsymbol{V}\in\mathbb{R}^{m\times d}$,$d$是特征维度.

3.2　多模态特征融合

现有MNER方法大多通过跨模态注意力、共注意力等机制实现多模态特征融合,然而,这些机制存在注意力偏置问题,即过度关注某些特定输入,而忽略其他重要信息,并且计算资源消耗较大. 考虑到本方法将图像描述为自然语言,消除了模态间差异,故利用BERT强大的上下文理解能力,联合编码图像描述语句和文本,以更好地捕获图像和文本之间的关联关系,更有效地整合图像描述语句和输入文本的语义特征. 图文联合编码的目标是捕获多模态之间的互补性,它能够同时编码多个模态的信息,获得更丰富、更全面的语义表示.

具体来说,将视觉嵌入$\boldsymbol{V}$作为文本嵌入$\boldsymbol{S}$的前缀,拼接后输入BERT中,利用其中的自注意力机制来捕获融合了视觉信息的多模态表征$\boldsymbol{C}$,其形式化表达如式(1)所示:

$$\boldsymbol{C}=\mathrm{BERT}(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m;\ \boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n)\tag{1}$$

在输出的$m+n$个特征向量中,选取与文本长度一

致的后 n 个向量作为多模态表征 $\boldsymbol{C}=\left[\boldsymbol{c}_1^s,\boldsymbol{c}_2^s,\cdots,\boldsymbol{c}_n^s\right]$,其中,$\boldsymbol{C}\in\mathbb{R}^{n\times d}$,$d$ 是特征维度.

4 标签对齐与融合

CLGLF方法使用KL散度损失函数对齐多模态表征和文本表征的预测向量,然后调用CRF对两个预测向量分别解码得到候选标签和文本标签. 在此基础上,该方法基于置信学习[27](Confident Learning,CL)的思想,提出了一个置信阈值引导标签融合(Confidence Threshold Guides Label Fusion,CTGLF)策略,通过置信分数和置信阈值共同指导MNER任务的最终预测.

4.1 标签对齐与解码

在本领域中,可以通过对比学习[15]或计算相似度得分[28]实现对齐,前一种方法过度依赖负样本和损失函数的选择,致使对齐效果不稳定. 后者,相似度计算是一种对称性的度量,难以捕捉两个特征之间的微小差异,导致对齐结果不准确. 因此,本方法利用KL散度损失函数实现标签对齐,KL散度的非对称性能更精确地量化两个概率分布之间的差异. 通过最小化KL散度,可以使多模态表征预测向量和文本表征预测向量的表示更一致,从而达到对齐标签的目的. 具体的,对于多模态表征 $\boldsymbol{C}$ 和文本表征 $\boldsymbol{X}$,使用同一个全连接层分别将其投影到命名实体识别任务的预测空间中,如式(2)和式(3)所示:

$$\boldsymbol{C}'=\mathrm{FC_p}(\boldsymbol{C}),\boldsymbol{C}'\in\mathbb{R}^{n\times 11} \tag{2}$$

$$\boldsymbol{X}'=\mathrm{FC_p}(\boldsymbol{X}),\boldsymbol{X}'\in\mathbb{R}^{n\times 11} \tag{3}$$

其中,$\boldsymbol{C}'$ 为多模态表征预测向量,$\boldsymbol{X}'$ 为文本表征预测向量,$\mathrm{FC_p}$ 为全连接层,n 为特征数量,11为投影的目标维度.

接下来,使用KL散度损失函数最小化多模态表征预测向量 $\boldsymbol{C}'$ 和文本表征预测向量 $\boldsymbol{X}'$ 之间的差距,以确保它们具有一致的语义表示,从而减轻视觉噪声对模型的影响,同时实现候选标签和文本标签之间的初步融合. KL散度损失函数的计算如式(4)所示:

$$L_{\mathrm{KL}}=\sum_{y\in y(x)}\mathrm{P}(y|\boldsymbol{C}')\log\mathrm{P}(y|\boldsymbol{X}') \tag{4}$$

其中,$y(x)$ 为所有可能的预测向量的集合.

然后,将 $\boldsymbol{C}'$ 和 $\boldsymbol{X}'$ 分别输入标签解码器中,生成候选标签 $\boldsymbol{E}$ 和文本标签 $\boldsymbol{F}$. 本方法选取CRF作为标签解码器,它在序列标记任务中考虑了标签之间的依赖关系,从而产生更高精度的输出标签. 候选标签的概率分布计算如式(5)和式(6)所示:

$$P\left(\boldsymbol{E}|\boldsymbol{S},\boldsymbol{V}\right)=\frac{\exp\left(\mathrm{score}(\boldsymbol{C}',\boldsymbol{E})\right)}{\sum_{\boldsymbol{E}'}\exp\left(\mathrm{score}\left(\boldsymbol{C}',\boldsymbol{E}'\right)\right)} \tag{5}$$

$$\mathrm{score}\left(\boldsymbol{C}',\boldsymbol{E}\right)=\sum_{i=0}^{n}\boldsymbol{A}_{e_i,e_{i+1}}+\sum_{i=1}^{n}\boldsymbol{H}_{c_i',e_i} \tag{6}$$

其中,$\mathrm{score}(\boldsymbol{C}',\boldsymbol{E})$ 为序列标注分数,由转移分数和发射分数两部分组成,$\boldsymbol{A}_{e_i,e_{i+1}}$ 为标签 e_i 到标签 e_{i+1} 的转移分数,$\boldsymbol{H}_{c_i',e_i}$ 为将 c_i' 预测为标签 e_i 的概率.

依据上述处理过程,计算文本标签的概率分布 $P\left(\boldsymbol{F}|\boldsymbol{S}\right)$. 本文将对数似然损失作为候选标签生成任务,以及文本标签生成任务的损失函数,如式(7)和式(8)所示.

$$L_{\mathrm{MNER}}=-\sum_{i=1}^{n}(\log P(e_i^*|\boldsymbol{C}')) \tag{7}$$

$$L_{\mathrm{NER}}=-\sum_{i=1}^{n}(\log P(f_i^*|\boldsymbol{X}')) \tag{8}$$

其中,e_i^* 和 f_i^* 为训练批次中的第 i 个样本.

4.2 置信阈值引导标签融合

KL散度损失函数从语义层面对齐两组标签,忽略了标签层面的融合. 事实上,基于纯文本的命名实体识别可以产生正确的标签. 因此,本文提出了一个置信阈值引导的标签融合策略,指导候选标签和文本标签融合,有效地减少了多模态语义偏差. 对于标签融合,一些方法侧重于以文本标签为主,如转换矩阵[21]中的元素是文本标签到多模态候选标签的转换概率,但通常情况下,候选标签的准确率高于文本标签,而本文所提的置信阈值引导的标签融合策略正是以文本标签辅助筛选候选标签,从而得到更精确的标签序列 $y=\{y_1,y_2,\cdots,y_n\}$. CTGLF的具体处理流程,如算法1所示.

算法1 置信阈值引导的标签融合算法

输入: 候选标签 $\boldsymbol{E}$,文本标签 $\boldsymbol{F}$,多模态表征 $\boldsymbol{C}$,置信阈值CT
输出: 预测标签 y
BEGIN
1: FOR i IN $\boldsymbol{c}_i^s$ DO
2: $\quad\boldsymbol{c}_i^{s'}\leftarrow\mathrm{FC}_z(\boldsymbol{c}_i^s)$
3: $\quad\mathrm{CS}_i\leftarrow\mathrm{sigmoid}(\boldsymbol{c}_i^{s'})$
4: $\quad$IF $\mathrm{CS}_i>\mathrm{CT}$
5: $\quad\quad y_i=f_i$
6: $\quad$ELSE
7: $\quad\quad y_i=e_i$
8: END FOR
END

对多模态特征 $\boldsymbol{C}$ 执行一个全连接层(见第1行和第2行),然后使用sigmoid激活函数进行处理(见第3行),可以得到每个特征的置信分数(Confidence Score,CS),它准确度量了多模态表征 $\boldsymbol{C}$ 的质量. 置信分数 z_i 的计算如式(9)所示:

$$\boldsymbol{Z}=\sigma(\mathrm{FC}_z(\boldsymbol{C})) \tag{9}$$

其中,$\boldsymbol{Z}=\left[z_1,z_2,\cdots,z_n\right]$,$\boldsymbol{Z}\in\mathbb{R}^{n\times 1}$,$\sigma$ 为sigmoid激活函数,FC_z 为全连接层.

本方法使用网格搜索,在[0, 1]范围内,以0.05为步长,循环遍历搜索网格点来确定置信阈值(Confidence Threshold,CT)的最优取值,结果表明,当CT取0.8时,方法的效果最佳,故本文置信阈值设置为0.8. 通过将所设定的置信阈值与置信分数z_i进行比较,可以筛选出候选标签中的错误预测. 若候选标签的置信分数高于该阈值,则使用单模态模型生成的文本标签替代候选标签(见第4行和第5行),否则保留候选标签(见第6行和第7行),最终更准确的结果标签y,如式(10)所示:

$$y_i=\begin{cases} f_i, z_i > \mathrm{CT} \\ e_i, z_i \leqslant \mathrm{CT} \end{cases} \tag{10}$$

其中,$y_i \in Y$,Y是具有标准BIO格式的预定义标签集.

在模型训练过程中,总体损失函数L是候选标签生成任务损失L_{MNER}、文本标签生成任务损失L_{NER}和KL散度损失L_{KL}的线性组合,如式(11)所示:

$$L=L_{\mathrm{MNER}}+L_{\mathrm{NER}}+L_{\mathrm{KL}} \tag{11}$$

5　方法验证及结果分析

5.1　实验设计

为验证本文提出的方法,在Twitter-2015和Twitter-2017多模态数据集上进行实验,使用召回率(Recall,R)、精确率(Precision,P)和F_1值(F_1-score,F_1)作为评价指标来评估所提方法的有效性.

Twitter-2015和Twitter-2017的统计数据如表1所示.

方法模型中各参数的默认值为:epoch=30、Batch-Size=32、lr=3×10^{-5},文本输入和图像描述的最大长度分别设置为80和20,文本特征表示使用BERT预训练模型,图像特征表示使用BLIP-2预训练模型,预训练模型的参数均保持原始设置.

5.2　对比实验

为验证本文方法的有效性,搭建实验环境,依据实验设计开展实验,并将CLGLF方法的实验结果与MSB[4]、RpBERT[8]、UMGF[10]、UMT[21]等7种相关的主流方法进行对比,其结果如表2所示.

表1　多模态数据集数据统计表

实体类别	Twitter-2015			Twitter-2017		
	训练集	验证集	测试集	训练集	验证集	测试集
PER	2 217	552	1 816	2 943	626	621
LOC	2091	552	1 697	731	173	178
ORG	928	247	839	1 674	375	395
MISC	940	225	726	701	150	157
图文对	4 000	1 000	3 257	3 373	723	723

由表2可以看出,CLGLF在两个数据集上的F_1值,分别比使用图像整体编码的MSB高出1.62%、3.08%,比使用ResNet提取区域视觉特征的RpBERT高出0.69%,比使用细粒度对象级视觉特征的UMGF高出0.24%、1.89%. 上述实验结果表明,本文使用BLIP-2生成的图像描述对方法性能产生了积极作用.

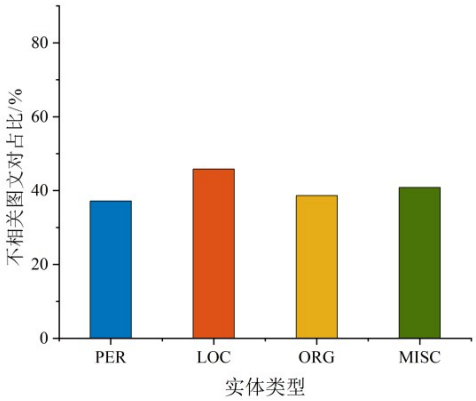
此外,本文方法还优于其他减少多模态语义偏差的模型,在两个数据集上获得最优F_1值. CLGLF与在模态融合过程中减少多模态语义偏差的HvpNet和QG相比,Twitter-2015上的F_1值分别提高了0.76%、0.39%,Twitter-2017上分别提高了1.60%、0.46%,这是因为本文方法从语义和标签两个层面上共同减少多模态语义偏差,从而产生了更准确的标签序列;CLGLF与在解码阶段减少多模态语义偏差的UMT和M3S相比,Twitter-2015上F_1值的差距达到1.68%、0.06%,Twitter-2017上的差距达到2.09%、1.34%,其中一个重要的原因是M3S未充分利用NER生成的标签,以筛选有偏差的MNER标签,只是简单相加了两者的损失来训练模型,而CLGLF则在缩小模态差异的基础上,提出了一个置信阈值引导的标签融合策略,利用NER标签辅助生成结果标签,更好地减少了在解码阶段的多模态语义偏差.

表2　对比实验结果分析表　　单位:%

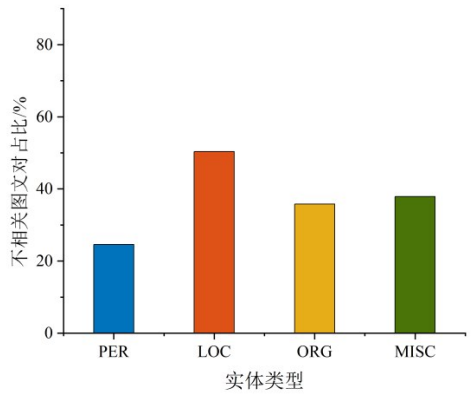
模型	Twitter-2015							Twitter-2017						
	Single Type(F_1)				Overall			Single Type(F_1)				Overall		
	PER	LOC	ORG	MISC	P	R	F_1	PER	LOC	ORG	MISC	P	R	F_1
MSB[4](2020)	**86.44**	77.16	52.91	36.05	74.97	72.04	73.47	—	—	—	—	85.20	83.60	84.32
RpBERT[8](2021)	—	—	—	—	—	—	74.40	—	—	—	—	—	—	—
UMGF[10](2021)	84.26	**83.17**	62.45	**42.42**	74.49	75.21	74.85	91.92	85.22	83.13	69.83	86.54	84.50	85.51
HvpNet[14](2022)	85.74	81.78	61.92	40.81	73.06	**75.65**	74.33	92.28	84.81	84.37	65.20	85.30	86.31	85.80
MNER-QG[19](2023)	85.31	81.65	63.41	41.32	**77.43**	72.15	74.70	92.92	**86.19**	84.52	**71.67**	**88.26**	85.65	86.94
UMT[21](2020)	85.24	81.58	63.03	39.45	71.67	75.23	73.41	91.56	84.73	82.24	70.10	85.28	85.34	85.31
M3S[23](2023)	86.05	81.32	62.97	41.36	74.92	75.14	75.03	92.73	84.81	82.49	69.53	86.93	85.21	86.06
CLGLF (ours)	86.30	81.92	**63.59**	38.52	75.21	74.98	**75.09**	**93.59**	85.87	**84.87**	70.06	86.45	**88.37**	**87.40**

注:表中数据是通过运行原论文提供的源代码或引用原论文中的实验结果得到的,“—”意味着相应的结果未显示.

CLGLF在Twitter-2015和Twitter-2017数据集中的单项指标LOC和MISC上均表现不佳，针对该情况，本文对数据集中的内容进行分析，通过计算文本和图像描述之间的余弦相似度，以评估各个实体类型中图文间的关联程度，结果如图2所示. 从图2中可以看出，在Twitter-2015数据集的LOC和MISC类别中，图文不相关的数据分别约占45.82%、40.86%，在Twitter-2017数据集中约占50.35%、37.89%，且Twitter-2017数据集整体的图文相关程度高于Twitter-2015. 本文方法主要试图降低多模态语义偏差，即处理图文弱关联情况对MNER的影响，但对图文显著无关的情况有待改进. 因此，CLGLF方法在LOC和MISC此类图文无关占比较大的指标上表现欠佳.



(*a*) Twitter-2015数据集中不相关图文数据占比



(*b*) Twitter-2017数据集中不相关图文数据占比

图2 Twitter数据集中不相关图文数据占比

在Twitter-2015和Twitter-2017数据集中，CLGLF的总体精确率(P)均低于MNER-QG，其原因是，MNER-QG通过机器阅读理解框架，明确地将高度相关的文本实体与相应的视觉区域对齐，从而实现高精确率(P). 针对当前CLGLF精确率较低的问题，在下一步的工作中，将尝试通过细粒度的图文匹配策略，实现文本和图像之间更精准的对齐，提高方法的准确率.

5.3 消融实验

为了验证CLGLF模型中核心组件的有效性，进行了消融实验，结果如表3所示.

表3中，“w/o KL”表示删除KL散度引导标签对齐模块，在两个数据集中，w/o KL的总体F_1值较CLGLF分别下降了2.73%，1.97%. 分析可知，KL散度损失函数在一定程度上提升了模型的精度；“w/o CTGLF”表示删除整个置信阈值引导的标签融合模块，在两个数据集中，w/o CTGLF的总体F_1值较CLGLF分别下降了0.20%、0.72%. 分析可知，CTGLF模块的确在KL散度对齐标签的基础上进一步提高了结果标签的精确度.

表3 消融实验结果表 单位：%

方法	Twitter-2015	Twitter-2017
CLGLF (ours)	75.09	87.40
w/o KL	72.63(↓2.73)	85.43(↓1.97)
w/o CTGLF	74.89(↓0.20)	86.68(↓0.72)
rep.CT	74.07(↓1.02)	85.92(↓1.48)
rep.CS	74.78(↓0.31)	86.94(↓0.46)

为了证明CTGLF模块内部组件的不可替代性，分别对CT和CS执行替换操作. “rep.CT”表示替换CTGLF模块中置信阈值的设置，具体为，对辅助任务中的文本表征执行置信分数的生成操作，并与多模态表征的置信分数逐一比较以确定标签的取舍. 在两个数据集中，“rep.CS”的总体F_1值较CLGLF分别下降了1.02%、1.48%. 分析可知，设置置信阈值不仅使方法效果更好，而且模型结构更简洁；“rep.CS”表示替换CTGLF模块中的置信分数，具体为，对辅助任务中的文本表征执行置信分数的生成操作，并将生成的置信分数与置信阈值进行比较，以筛选文本标签，并使用候选标签替换，实现标签融合. 在两个数据集中，“rep.CS”的总体F_1值较CLGLF分别下降了0.31%、0.46%. 分析可知，相较于NER，引入图像的MNER准确性更高，故应该以MNER产生的预测标签序列为主体，NER作为辅助任务进行校正，反之，则性能下降.

上述实验结果证明，KL散度损失函数、置信阈值引导的标签融合模块及其内部组件都是模型不可或缺的.

5.4 视觉编码器的影响

本小节验证视觉编码器BLIP-2对方法性能的影响，实验结果如表4所示.

分别用生成图像描述的CNN-RNN[29]和生成对象标签的Mask RCNN[30]替换本文方法使用的BLIP-2，实验结果表明，这两种图像编码方式的总体F_1值均有不同程度的下降，在两个数据集中，最大降幅分别达到0.32%和0.84%，这是因为传统的图像描述和对象标签生成技术，虽然缩小了两种模态之间的语义差异，但不

表4　视觉编码器消融结果表　　单位:%

方法	视觉编码器	Twitter-2015	Twitter-2017
CLGLF	BLIP-2	75.09	87.40
	CNN-RNN	74.77(↓0.32)	86.56(↓0.84)
	Mask RCNN	74.83(↓0.26)	86.90(↓0.50)
MSB	InceptionV3	73.47	84.32
	BLIP-2	73.56(↑0.09)	85.56(↑1.24)

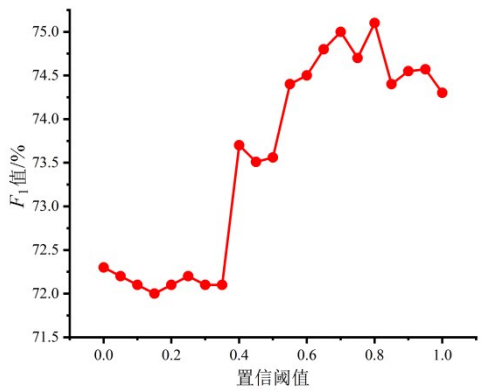
能准确地表述图像语义,更多的是凭空推测图像中的对象及其之间的关系. 相比之下,BLIP-2强大的视觉语义理解能力,能够捕获图像中的关键特征和复杂的语义内容,以产生更精确、流畅的图像描述语句,从而较好地解决模态间语义差异问题,利于多模态特征融合. 此外,BLIP-2还具备良好的泛化能力,可以处理多种类型的图像数据.

本文的实验还用BLIP-2替换了MSB模型中的InceptionV3视觉编码器,替换后的MSB在两个数据集上的性能均有所提升. 综合实验结果可知,本文所调用的预训练模型BLIP-2对MNER方法性能的提升是有效的.

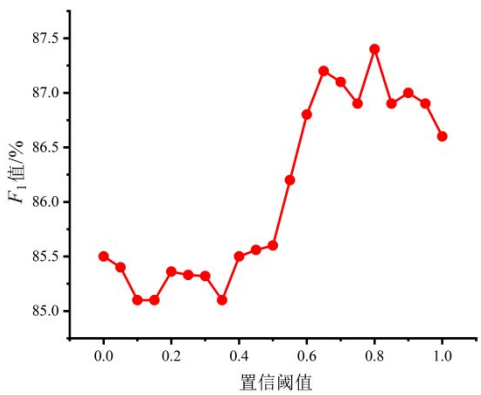
5.5　参数敏感性分析

置信阈值(CT)是CLGLF方法的关键超参数,本文使用网格搜索法寻找最优置信阈值,以实现方法的最佳性能,同时分析了CT的不同设置对方法性能的影响. 网格搜索(grid search)是一种常用的寻优调参方法,它通过遍历超参数的所有可能组合,使用交叉验证来评估每种参数组合的性能,从而选择最优超参数设置. 在本方法中,将置信阈值的取值范围设定为[0,1],并以0.05为步长,确定一个20×1的搜索网格,然后循环遍历所有网格点,在训练过程中记录每个候选阈值对应的F_1值,并进行比较. 最终确定使F_1值最优的置信阈值为0.8. 值得注意的是,本文使用的Twitter数据集是由大部分内容相关的图文对构成,然而,不同多模态数据集之间的数据分布及内容存在差异性,因此,若要在其他多模态数据集上测试该模型,则需要重新选择适用于其数据特点的置信阈值,以获得最佳性能. 不同CT取值对应的F_1值如图3所示,其中,图3(a)为Twitter-2015数据集上的结果,图3(b)为Twitter-2017数据集上的结果.

正如预期的一样,在两个数据集上,CT与F_1值的分布可大致拟合为一个开口向下的抛物线曲线. 具体来说,当CT设置为0.0时,模型退化为传统基于纯文本的命名实体识别,随着CT的逐渐增长,方法的性能呈上升趋势,这是因为候选标签在逐渐发挥主体作用. 通常来说视觉信息有助于文本实体及其类型的识别,特别是当文本语义有歧义时,但并不总是如此,当文本和图像的语义关系不匹配时,不相关或具有误导性的图像信息可能会引入噪声,导致候选标签存在误差或错误,映射到图3中,CT在[0.05,0.35]范围内的实验数据验证了该结论. 当CT的值增长至1.0时,模型最终的标签都来自于候选标签,没有文本标签的辅助,此时方法的性能不是最优. 当CT设置为0.8时,本文模型在两个数据集上均获得最佳的F_1值. 以上实验结果表明,有NER辅助的多模态命名实体识别方法可以获得更好的预测结果.



(a) Twitter-2015数据集中置信阈值与F_1值的关系



(b) Twitter-2017数据集中置信阈值与F_1值的关系

图3　Twitter数据集中置信阈值与F_1值的关系

6　总结

本文的置信学习引导标签融合的多模态命名实体识别方法,使用了BLIP-2预训练模型生成图像描述,在缩小模态差异的同时,优化了视觉语义. 为了减少多模态语义偏差,构造了一个置信阈值引导的标签融合机制,通过置信阈值筛选有偏差的候选标签,并用相应位置的文本标签进行替换,实现了标签融合. 在Twitter-2015和Twitter-2017两个公共数据集上的对比实验,分别从图像语义信息挖掘和多模态特征融合两个方面,对方法的有效性进行了验证. 然而,该方法也存在着一定的局限性,特别是当数据集中有较多图文无关的数据时,方法的效果受限. 在接下来的研究中,可尝试通

过与其他相关方向[31]做联合研究来抑制无关图像对文本实体识别的消极影响,进一步提升方法模型的健壮性.

参考文献

[1] 张聿远, 闫文君, 张立民. 基于多模态特征融合网络的空时分组码识别算法[J]. 电子学报, 2023, 51(2): 489-498.
ZHANG Y Y, YAN W J, ZHANG L M. Space-time block code recognition algorithm based on multi-modality features fusion network[J]. Acta Electronica Sinica, 2023, 51(2): 489-498. (in Chinese)

[2] MOON S, NEVES L, CARVALHO V. Multimodal named entity recognition for short social media posts[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 852-860.

[3] LU D, NEVES L, CARVALHO V, et al. Visual attention model for name tagging in multimodal social media[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 1990-1999.

[4] ASGARI-CHENAGHLU M, FEIZI-DERAKHSHI M R, FARZINVASH L, et al. CWI: A multimodal deep learning approach for named entity recognition from social media using character, word and image features[J]. Neural Computing and Applications, 2022, 34(3): 1905-1922.

[5] ZHANG Q, FU J L, LIU X Y, et al. Adaptive co-attention network for named entity recognition in tweets[C]//Proceedings of the Thirty-Second AAAI Conferenceon Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans: AAAI Press, 2018: 5674-5681.

[6] WANG X W, YE J B, LI Z X, et al. CAT-MNER: Multimodal named entity recognition with knowledge-refined cross-modal attention[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2022: 1-6.

[7] 李晓腾, 张盼盼, 勾智楠, 等. 基于多任务学习的多模态命名实体识别方法[J]. 计算机工程, 2023, 49(4): 114-119.
LI X T, ZHANG P P, GOU Z N, et al. Multi-modal named entity recognition method based on multi-task learning[J]. Computer Engineering, 2023, 49(4): 114-119. (in Chinese)

[8] SUN L, WANG J Q, ZHANG K, et al. RpBERT: A text-image relation propagation-based BERT model for multimodal NER[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13860-13868.

[9] ZHENG C M, WU Z W, WANG T, et al. Object-aware multimodal named entity recognition in social media posts with adversarial learning[J]. IEEE Transactions on Multimedia, 2021, 23: 2520-2532.

[10] ZHANG D, WEI S Z, LI S S, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14347-14355.

[11] WU Z W, ZHENG C M, CAI Y, et al. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 1038-1046.

[12] 钟维幸, 王海荣, 王栋, 等. 多模态语义协同交互的图文联合命名实体识别方法[J]. 广西科学, 2022, 29(4): 681-690.
ZHONG W X, WANG H R, WANG D, et al. Image-text joint named entity recognition method based on multimodal semantic interaction[J]. Guangxi Sciences, 2022, 29(4): 681-690. (in Chinese)

[13] WANG X Y, GUI M, JIANG Y, et al. ITA: Image-text alignments for multi-modal named entity recognition[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: 3176-3189.

[14] CHEN X, ZHANG N Y, LI L, et al. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction[C]//Findings of the Association for Computational Linguistics: NAACL 2022. Stroudsburg: Association for Computational Linguistics, 2022: 1607-1618.

[15] XU B, HUANG S Z, SHA C F, et al. MAF: A general matching and alignment framework for multimodal named entity recognition[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. New York: ACM, 2022: 1215-1223.

[16] ZHANG Z X, MAI W X, XIONG H L, et al. A token-wise graph-based framework for multimodal named entity recognition[C]//IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2023: 2153-2158.

[17] MAI W X, ZHANG Z X, LI K T, et al. Dynamic graph construction framework for multimodal named entity recognition in social media[J]. IEEE Transactions on Computational Social Systems, 2024, 11(2): 2513-2522.

[18] ZHAO F, LI C, WU Z, et al. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal NER[C]//Proceedings of the 30th

ACM International Conference on Multimedia. New York: ACM, 2022: 3983-3992.

[19] JIA M, SHEN L, SHEN X, et al. MNER-QG: An end-to-end MRC framework for multimodal named entity recognition with query grounding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(7): 8032-8040.

[20] ZHANG X, YUAN J L, LI L, et al. Reducing the bias of visual objects in multimodal named entity recognition[C]// Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. New York: ACM, 2023: 958-966.

[21] YU J F, JIANG J, YANG L, et al. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 3342-3352.

[22] LIU L P, WANG M L, ZHANG M Z, et al. UAMNer: Uncertainty-aware multimodal named entity recognition in social media posts[J]. Applied Intelligence, 2022, 52 (4): 4109-4125.

[23] WANG J, YANG Y, LIU K Y, et al. M3S: Scene graph driven multi-granularity multi-task learning for multi-modal NER[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 111-120.

[24] LI J N, LI D X, SAVARESE S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the 40th International Conference on Machine Learning (ICML'23), Honolulu: JMLR.org, 2023: 19730-19742.

[25] PAN Y H. On visual understanding[J]. Frontiers of Information Technology & Electronic Engineering, 2022, 23 (9): 1287-1289.

[26] ZHANG S S, ROLLER S, GOYAL N, et al. OPT: Open pre-trained transformer language models[EB/OL]. (2022-06-21)[2023-12-11]. http://arxiv.org/abs/2205.01068.

[27] NORTHCUTT C, JIANG L, CHUANG I. Confident learning: Estimating uncertainty in dataset labels[J]. Journal of Artificial Intelligence Research, 2021, 70: 1373-1411.

[28] CHENG J, LONG K F, ZHANG S, et al. Text-image scene graph fusion for multi-modal named entity recognition[J]. IEEE Transactions on Artificial Intelligence, 2023, PP(99): 1-12.

[29] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 3156-3164.

[30] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2980-2988.

[31] 杜晋华, 尹浩, 冯嵩. 中文电子病历命名实体识别的研究与进展[J]. 电子学报, 2022, 50(12): 3030-3053.
DU J H, YIN H , FENG S. Research and development of named entity recognition in Chinese electronic medical record[J]. Acta Electronica Sinica, 2022, 50(12): 3030-3053. (in Chinese)

作者简介

王海荣　女,1977年出生,宁夏回族自治区石嘴山市人. 现为北方民族大学计算机科学与工程学院教授. 主要研究方向为大数据知识工程与智能信息处理.
E-mail: bmdwhr@163.com

王　彤　女,1999年出生,山西省吕梁市人. 现为北方民族大学计算机科学与工程学院硕士研究生. 主要研究方向为多模态信息抽取.
E-mail: is_wangtong@163.com

徐　玺　男,1997年出生,云南省昭通市人. 北方民族大学计算机科学与工程学院硕士研究生. 主要研究方向为多模态信息抽取.
E-mail: 2533100032@qq.com

荆博祥　男,1999年出生,山西省运城市人. 现为北方民族大学计算机科学与工程学院硕士研究生. 主要研究方向为多模态知识推理.
E-mail: j1131220481@163.com

陈芳萍　女,1997年出生,甘肃省定西市人. 现为北方民族大学计算机科学与工程学院硕士研究生. 主要研究方向为多模态信息抽取.
E-mail: 2483757064@qq.com