文章编号:1671-6833(2024)02-0060-12

多模态命名实体识别方法研究进展

王海荣[1,2], 徐 玺[1], 王 彤[1], 荆博祥[1]

(1.北方民族大学 计算科学与工程学院,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:为了解决多模态命名实体识别(MNER)研究中存在的文本特征语义不足、视觉特征语义缺失、图文特征融合困难等问题,多模态命名实体识别方法相继被提出。首先,总结了多模态命名实体识别方法的整体框架以及各部分常用的技术,随后对其进行梳理并分类为基于 BiLSTM 的 MNER 方法和基于 Transformer 的 MNER 方法,并根据模型结构将其划分为前融合模型、后融合模型、Transformer 单任务模型、Transformer 多任务模型等 4 类模型结构。其次,在 Twitter-2015、Twitter-2017 2 个数据集上,分别对这 2 类方法进行实验,结果表明:多特征协同表示能增强各模态特征的语义,多任务学习能够促进模态特征融合或者结果融合,从而提升 MNER 的准确性。建议在 MNER 的未来研究中,着重关注通过多特征协同表示来增强模态语义,通过多任务学习促进模态特征融合或结果融合等方向的研究。

关键词:多模态命名实体识别; Transformer; BiLSTM; 多模态融合; 多任务学习

中图分类号:TP301.6;TP391.1 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2024.02.001

命名实体识别任务是从数据中识别出专有名词,最早在信息理解会议[1]上被提出。随后形成了基于规则[2]和词典[3]的方法、机器学习的方法、深度学习的方法等 3 类命名实体识别方法。基于规则和词典的方法通过字符匹配进行信息抽取,适用于数据更新较少的领域,但规则和词典制定成本较高。基于机器学习的方法将命名实体识别任务视为分类问题,并提出了如 HMM-based[4]、CRF-based[5]的模型,该方法减少了人工成本,但选取特征的质量决定了算法的性能。基于深度学习的方法具有自动挖掘高质量上下文特征的能力,研究者相继提出了 CNN-based[6]、BiLSTM-based[7]、Transformer-based[8-10]、GNN-based[11-13]等模型,但要求文本有充足的上下文特征,因此在长文本数据集中的性能表现更好,在短文本数据集中性能表现不佳。

传统的文本语义增强主要依赖字符特征[14]、词汇信息[15]、知识图谱[16-17]、检索[18]、标签信息[19]等外部文本数据,也结合了多任务学习来增强命名实体识别的能力。王蓬辉等[20]采用基于生成对抗的数据增强算法来解决标注数据不足的问题。余传明等[21]提出了实体和事件联合抽取模型,从而在 2 个任务中均取得了更好的效果。武国亮等[22]提出将命名实体识别任务的输出反馈到输入端,来解决多任务联合学习产生的损失不平衡问题。但随着社交媒体平台的广泛应用,以文本、图像为主要媒介的多模态数据快速增长,为了从这些多模态数据中挖掘语义,进而增强文本特征,人们提出了多模态命名实体识别(multimodal named entity recognition,MNER)方法。MNER 研究难点是如何融合多模态特征中有益信息,并过滤有害信息。早期研究[23-25]关注使用视觉特征增强静态词表示的方法,取得了一些研究成果。范涛等[26]将 MNER 迁移到了地方志领域的实体识别研究。近年来,随着预训练语言模型的发展,MNER 方法的研究重点逐步转向采用 Transforme 融合特征,取得了新的研究成果。现有的 MNER 方法可分为 4 类,如表 1 所示。

收稿日期:2023-08-24;**修订日期**:2023-10-28

基金项目:宁夏回族自治区自然科学基金资助项目(2023AAC03316);宁夏回族自治区教育厅高等学校科学研究重点项目(NYG2022051);北方民族大学中央高校基本科研业务费专项资金资助项目(2022PT_S04);北方民族大学校级科研项目(2021XYZJK06)

作者简介:王海荣(1977—),女,宁夏石嘴山人,北方民族大学副教授,博士,主要从事大数据知识工程与智能信息处理方面的研究,E-mail:bmdwhr@163.com。

引用本文:王海荣,徐玺,王彤,等.多模态命名实体识别方法研究进展[J].郑州大学学报(工学版),2024,45(2):60-71.(WANG H R,XU X,WANG T,et al. Research progress of multimodal named entity recognition [J]. Journal of Zhengzhou University(Engineering Science),2024,45(2):60-71.)

表 1　多模态命名实体识别方法分类表

Table 1　Multimodal named entity recognition method classification table

方法	典型模型	优点	缺点
前融合模型	MA[27]、VAM[28]	简单、直接的方式实现多模态融合,性能优于 NER	图文特征存在语义鸿沟
后融合模型	ACN[29]、GAN[30]	文本语义和图像语义差距较小;多模态特征质量较高	文本特征的语义低
Transformer 单任务模型	UMGF[31]、MAF[32]	文本语义和图像语义更接近;图像和文本特征深度融合	视觉特征语义存在偏差
Transformer 多任务模型	UMT[33]、ITA[34]	多任务协同优化多模态特征,多模态特征语义更准确	视觉特征缺失

2018 年,Moon 等[27]首次在 BiLSTM-CRF 模型中融入了视觉特征,提出了多模态实体识别方法,提出 MA[27]模型。VAM[28]、CWI[29]等模型也被提出。这些模型均使用注意力机制为文本表示和视觉特征分配权重,拼接得到多模态特征,再通过 BiLSTM+CRF 网络挖掘上下文特征并解码,将此类模型归纳为前融合模型。前融合模型中使用 Glove 表示单词,导致图像特征与文本特征间的语义存在巨大鸿沟。针对该问题,一些学者提出 ACN[29]、GAN[30]、DCN[35]等模型,先使用 BiLSTM 挖掘文本中上下文特征以增强单词的实体语义,然后采用注意力机制作为多模态融合层,得到多模态特征,将此类模型归纳为后融合模型。

为了进一步缩小文本与图像特征的语义差距,2020 年,基于 Transformer 的 MNER 方法首次被 Yu 等[33]提出,其中 Chen 等[36]使用 BERT 表示文本,并验证了提升单词语义的重要性,之后 UMGF[31]、MAF[32]、ITJ[37]、HSN[38]等模型相继被提出,这些模型堆叠多个 Transformer,对各模态特征进行编码、对齐或融合处理,得到多模态特征后,均只后接 1 个命名实体识别任务,本文将此类模型归纳为 Transformer 单任务模型。为了解决多模态特征与目标语义间的偏差问题,一是在文本表示和多模态表示上构建联合实体识别任务,以解决视觉偏差的问题,如 Yu 等[33]的边界检测任务,Wang 等[34]和 Liu 等[39]的文本视图命名实体识别任务。二是通过辅助任务联合训练多模态表示,增强特征的通用性。如李晓腾等[40]提出通过对比融合、实体聚类、边界检测等任务辅助学习多模态特征,Chen 等[41]结合关系抽取任务训练多模态特征,本文将此类模型归纳为 Transformer 多任务模型。上面所提的 4 类模型尚没有关注单视觉特征中图像语义丢弃的问题。

此外,Sui 等[42]构建文本和语音数据集并提出 M3T 模型,进一步验证多模态特征能帮助识别命名实体。Liu 等[43]提出使用合成的声学特征而不是真实的人类语音,并采用多头注意力机制融合文本和语音 2 种模态的特征,稳定地提高了中文命名实体识别的性能。冯皓楠等[44]提出了一种图文注意力融合的主题标签推荐的方法,并表明相比单模态输入,多模态方法具有更显著的优势。郑建兴等[45]提出了基于评论文本情感注意力的推荐方法,使用注意力机制聚合用户特征和项目特征信息,以得到联合嵌入,进而提升了模型的有效性。

1　MNER 方法框架

根据 MNER 各方法的特点,将 MNER 方法的框架划分为模态输入表示、上下文编码层、多模态融合层、标签解码和多任务融合层。多模态命名实体识别的基本框架如图 1 所示。

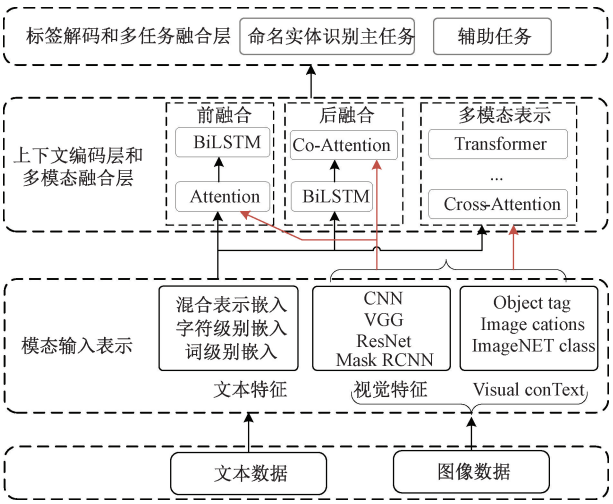

图 1　多模态命名实体识别的基本框架

Figure 1　Basic framework of MNER

模态输入表示是将各模态数据表示为特征向量。上下文编码器旨在挖掘特征的上下文依赖关系,主要采用 BiLSTM 模型。多模态融合层旨在融合多模态特征,主要采用 Attention、Co-Attention、Transformer、Transformer with Cross-Attention 模型。多模态融合层和上下文编码器呈现 3 种结构,前融合模型将图像信息传递给每个单词,再挖掘单词间的上下文特征;后融合模型则相反;多模态表示使用 Transformer 作为联合编码模型,融合多模态特征。标签解码层将多模态表示作为输入特征来预测标签。此外,当 MNER 方法结合了如对抗学习、边界

检测、关系抽取等辅助任务共同训练,将之归纳为多任务融合层。综上,可划分为 4 种模型结构:前融合模型、后融合模型、Transformer 单任务模型及 Transformer 多任务模型。

1.1 模态输入表示

文本模态输入表示主要采用字符嵌入、词嵌入、混合嵌入表示文本特征。字符嵌入 $\boldsymbol{X}^{\mathrm{c}}$ 通过 CNN 或 RNN 模型进行表示,缓解 out-of-vocabulary 的问题;词嵌入模型包括 CBOW[46]、Word2Vec[47]、Glove[48]、FastText[49]、ELMo[50]、BERT 等。对于输入的句子 $\boldsymbol{S}$,静态词向量可表示为

$$\boldsymbol{X}^{\mathrm{s}} = \mathrm{Glove}(\boldsymbol{S})。\tag{1}$$

动态词向量可表示为

$$\boldsymbol{X}^{\mathrm{b}} = \mathrm{BERT}(\boldsymbol{S})。\tag{2}$$

混合嵌入[51]可表示为 $\boldsymbol{X} = [\boldsymbol{X}^{\mathrm{s}};\boldsymbol{X}^{\mathrm{c}}]$。

对于输入的图像特征 $\boldsymbol{I}$,视觉模态输入表示采用预训练数据模型进行特征表示,区域视觉特征使用 ResNet[52]模型提取,可表示为

$$\boldsymbol{V}^{\mathrm{r}} = \mathrm{ResNet}(\boldsymbol{I})\ ,\boldsymbol{V}^{\mathrm{r}} \in \mathbf{R}^{n\times d}。\tag{3}$$

式中:d 为特征维度;n 为特征的数量。

层级视觉特征是提取视觉模型中的中间层特征,$\boldsymbol{V}^{\mathrm{g}'} = \{\boldsymbol{V}_0^{\mathrm{g}'},\boldsymbol{V}_1^{\mathrm{g}'},\cdots,\boldsymbol{V}_{c'}^{\mathrm{g}'}\}$,$c'$ 为编码层的数量,再利用线性层 $\mathrm{MLP}_i(\ )$ 将每个特征映射到统一空间,计算如下:

$$\boldsymbol{V}_i^{\mathrm{g}} = \mathrm{MLP}_i(\boldsymbol{V}_i^{\mathrm{g}'})\ ,i = 1,2,\cdots,c'。\tag{4}$$

利用 Mask RCNN[53]提取对象级视觉标签,可表示为

$$\boldsymbol{V}^{\mathrm{l}} = \mathrm{MaskRCNN}(\boldsymbol{I})。\tag{5}$$

图像标题使用图像字幕[54](image captioning,IC)提取,可表示为

$$\boldsymbol{V}^{\mathrm{cap}} = \mathrm{IC}(\boldsymbol{I})。\tag{6}$$

1.2 上下文编码层和多模态融合层

双向长短时记忆网络(bi-directional long-short term memory,BiLSTM)作为上下文编码器时,能提取单词上下文特征。自注意力机制能增强关键特征的权重。self-Attention(SA)、Multi Head self-Attention(MHSA),Cross-Attention(CA)、计算原理表示如下:

$$\mathrm{SA}(\ ) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}}{\sqrt{d}}\right)\cdot\boldsymbol{V};\tag{7}$$

$$\mathrm{MHSA}(\ ) = \boldsymbol{W}'[\mathrm{SA}_0(\ ),\mathrm{SA}_1(\ ),\cdots,\mathrm{SA}_{m-1}(\ )];\tag{8}$$

$$\mathrm{CA}(\ ) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_1\cdot\boldsymbol{K}_2}{\sqrt{d}}\right)\cdot\boldsymbol{V}_2。\tag{9}$$

式中:$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 为输入特征的投影向量;d 为 $\boldsymbol{Q}$ 的特征维度;m 为多头注意力的头数;$\boldsymbol{W}'$ 代表投影矩阵;$\boldsymbol{Q}_1$ 代表文本模态的输入特征的投影向量;$\boldsymbol{K}_2$、$\boldsymbol{V}_2$ 代表视觉模态的输入特征的投影向量。此外,multi head cross-attention(MHCA)是将 MHSA( )中的 SA( )替换为 CA( )。

Transformer 能获取到长距离依赖关系,由多个编码器堆叠形成,每个编码器由多头自注意力机制、前馈层及规范化层组成。

Transformer 作为多模态融合层时,其多模态特征融合技术主要为以下 3 种构建方式:①将文本表示和视觉表示投影到同一离散空间进行对齐;②使用视觉语言模型对文本和图像进行联合表示;③将视觉特征转化为自然语言描述,使用语言模型统一表示。

1.3 标签解码和多任务融合层

通常使用条件随机场(conditional random field,CRF)作为标签解码层,对多模态表示进行解码。设 $\boldsymbol{X} = \{\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ 为 CRF 的输入特征,$\boldsymbol{y} = \{\boldsymbol{y}_0,\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\}$,解码表示如下:

$$p(\boldsymbol{y}\mid\boldsymbol{X}) = \frac{\prod_{i=1}^{n}\phi(\boldsymbol{x}_i,\boldsymbol{y}_i,\boldsymbol{y}_{i-1})}{\sum_{\boldsymbol{y}'\in\boldsymbol{Y}}\prod_{i=1}^{n}\phi(\boldsymbol{x}_i,\boldsymbol{y}'_i,\boldsymbol{y}'_{i-1})}。\tag{10}$$

式中:$\boldsymbol{y}$ 为可能的标签序列的集合;势函数 $\phi(\boldsymbol{x}_i,\boldsymbol{y}_i,\boldsymbol{y}_{i-1}) = \exp(\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{x}_i + \boldsymbol{y}_{i-1}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{y}_i)$,$\boldsymbol{E}\in\mathbf{R}^{d\times u}$,$\boldsymbol{F}\in\mathbf{R}^{u\times u}$;$d$、$u$ 均为维度。

使用最大似然函数作为损失函数,计算如下:

$$\mathcal{L}(p(\boldsymbol{y}\mid\boldsymbol{X})) = \sum\log p(\boldsymbol{y}\mid\boldsymbol{X})。\tag{11}$$

$\boldsymbol{y}_{\mathrm{o}}$ 为预测输出序列得分最高的序列,计算如下:

$$\boldsymbol{y}_{\mathrm{o}} = \mathrm{argmax}_{\boldsymbol{y}}\,p(\boldsymbol{y}\mid\boldsymbol{X}),\tag{12}$$

多任务融合层中利用任务间的信息共享来训练模型参数,以全局最优的多模态特征或预测结果提升实体识别性能,增强模型的可用性,包括命名实体识别主任务和实现标签融合或优化多模态表征的辅助任务。

2 基于 BiLSTM 的 MNER 方法

基于 BiLSTM 的 MNER 方法以 BiLSTM 和 CRF 作为基础模块,并引入多模态融合层,实现文本和图像特征融合,以解决上下文特征匮乏的问题。根据多模态融合方法划分为前融合模型和后融合模型,并对各方法进行实验验证及对比分析。

2.1 前融合模型

前融合模型首先对各模态表示进行拼接或加权拼接,接着使用 BiLSTM 挖掘上下文特征,最后将融

合表示输入 CRF 中预测标签。前融合模型框架如图 2 所示。

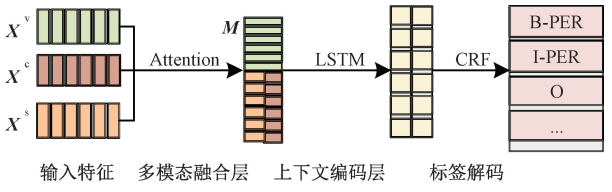

图 2　前融合模型

Figure 2　Pre-fusion model

Moon 等[27]的多模态融合层是先拼接单词表示、字符表示及区域视觉特征,将其映射到统一特征空间 $\boldsymbol{X}=\sigma([\boldsymbol{X}^s;\boldsymbol{X}^c;\boldsymbol{X}^v])$,$\boldsymbol{X}^v=\boldsymbol{V}^r$,$\sigma$ 为投影函数,使用注意力机制计算 $\boldsymbol{X}$ 中各模态权重 $[\boldsymbol{a}^s;\boldsymbol{a}^c;\boldsymbol{a}^v]$,得到融合表示 $\boldsymbol{M}$,以 $\boldsymbol{a}^s$ 为例,模态权重计算如下:

$$\boldsymbol{a}^s=\frac{\exp(\boldsymbol{a}^s)}{\exp(\boldsymbol{a}^s)+\exp(\boldsymbol{a}^c)+\exp(\boldsymbol{a}^v)};\quad(13)$$

$$\boldsymbol{M}=\boldsymbol{a}^s\boldsymbol{X}^s+\boldsymbol{a}^c\boldsymbol{X}^c+\boldsymbol{a}^v\boldsymbol{X}^v。\quad(14)$$

Lu 等[28]使用视觉注意力模型从图像中提取与文本最相关的图像特征,计算得到视觉上下文特征 $\boldsymbol{v}$,将 $\boldsymbol{v}$ 与词表示、字符表示拼接,得到融合表示 $\boldsymbol{M}=[\boldsymbol{v}_s;\boldsymbol{X}^s;\boldsymbol{X}^c]$,计算如下:

$$\boldsymbol{A}=\mathrm{softmax}(\boldsymbol{W}_1([\boldsymbol{X}^s;\boldsymbol{X}^c]\oplus\boldsymbol{X}^v));\quad(15)$$

$$\boldsymbol{v}=\sum\boldsymbol{a}_i\boldsymbol{X}_i^v,\boldsymbol{a}_i\in\boldsymbol{A},\boldsymbol{X}_i^v\in\boldsymbol{X}^v。\quad(16)$$

式中:$\boldsymbol{W}_1$ 为权值矩阵;$\oplus$ 为向量的求和函数;$\boldsymbol{A}$ 为视觉全局注意力权重。

Asgari-Chenaghlu 等[55]分别挖掘出字符特征、单词特征和图像特征的上下文特征并拼接这些上下文特征作为多模态融合表示。

经分析发现上述模型存在以下限制:单词表示的实体语义微弱。当单词的拼写错误,只能通过随机初始化进行表示,文本的实体语义被降低。此外,实现处于不同特征空间的图文特征对齐是很困难的。

2.2　后融合模型

后融合模型利用 BiLSTM 挖掘上下文特征,增强单词表示的实体语义,使用多模态注意力融合图文特征,再使用 CRF 模型解码。框架如图 3 所示。

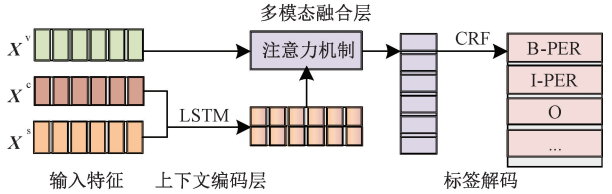

图 3　后融合模型

Figure 3　Pos-fusion model

Zhang 等[29]提出了共注意力网络(CoAttention),对于输入的图文特征,先计算基于文本上下文的文本特征 $\boldsymbol{h}^t=\mathrm{BiLSTM}([\boldsymbol{X}^s;\boldsymbol{X}^c])$,再分别通过共注意力机制计算基于文本注意力的视觉特征 $\boldsymbol{H}^{cv}$ 和基于视觉注意力的文本特征 $\boldsymbol{H}^{ct}$,通过门控机制 $\mathrm{gate}(\cdot)=\mathrm{softmax}(\mathrm{ReLU}(\cdot))$ 得到多模态表示 $\boldsymbol{M}$,计算如下:

$$[\boldsymbol{H}^{cv},\boldsymbol{H}^{ct}]=\mathrm{CoAttention}(\boldsymbol{X}^v,\boldsymbol{h}^t);\quad(17)$$

$$\boldsymbol{M}=\boldsymbol{h}^t+\boldsymbol{H}^{ct}\cdot\mathrm{gate}(\boldsymbol{H}^{cv}\oplus\boldsymbol{H}^{ct})。\quad(18)$$

但共注意力网络忽略了细粒度视觉对象和文本实体之间的关系,可能导致不同类型实体的错误预测。为此,Zheng 等[30]利用对抗学习优化投影函数将图文特征映射为一个共享的表示,然后采用双线性注意力计算每个单词和对象标签的细粒度语义关系,以及共同表示 $\boldsymbol{G}$,通过门控机制生成融合表示 $\boldsymbol{M}$,计算如下:

$$\boldsymbol{A}=\mathrm{softmax}(((\boldsymbol{one}\cdot\boldsymbol{P}^{\mathrm{T}})\circ\boldsymbol{h}^t\boldsymbol{W}_2)\boldsymbol{W}_3^{\mathrm{T}}\boldsymbol{X}^v);\quad(19)$$

$$\boldsymbol{G}=\boldsymbol{X}^v\cdot\boldsymbol{A}^{\mathrm{T}};\quad(20)$$

$$\boldsymbol{M}=\boldsymbol{\sigma}([\boldsymbol{G};\boldsymbol{h}^t])\circ\boldsymbol{G}+\boldsymbol{h}^t。\quad(21)$$

式中:$\boldsymbol{\sigma}()$ 为投影函数;$\boldsymbol{one}$ 为向量;$\boldsymbol{P}$ 为注意力得分的池化参数矩阵;$\boldsymbol{W}_2$、$\boldsymbol{W}_3$ 为参数矩阵;$\circ$为哈达玛积。

Wu 等[35]使用视觉标签表示图像语义,引入密集的共注意力机制建立单词和对象之间的关系,实现细粒度语义交互,得到多模态表示 $\boldsymbol{M}$。计算如下:

$$\boldsymbol{h}^{tsa}=\mathrm{SA}(\boldsymbol{h}^t,\boldsymbol{h}^t,\boldsymbol{h}^t);\quad(22)$$

$$\boldsymbol{h}^v=\mathrm{SA}(\boldsymbol{X}^v,\boldsymbol{X}^v,\boldsymbol{X}^v);\quad(23)$$

$$\boldsymbol{M}=\boldsymbol{h}^t+\mathrm{CA}(\boldsymbol{h}^{tsa},\boldsymbol{h}^v,\boldsymbol{h}^v)。\quad(24)$$

式中:$\boldsymbol{X}^v$、$\boldsymbol{h}^t$ 代表视觉特征、文本特征;SA()和 CA()分别代表自注意力机制和跨模态注意力机制。

2.3　方法分析

在 Twitter 2015 和 Twitter 2017 数据集上进行实验,使用评估指标[56]如召回率 R 和 $F1$ 值对 MNER 方法的有效性进行对比分析。本文选择 Lu 等[28]、Zhang 等[29]和 Chen 等[36]提出的基线模型,前融合模型选取 MA 和 VAM 模型,后融合模型选取增加字符表示的 ACN 模型,以及在 ACN 模型上融合其他方法的模型,即使用视觉区域特征的 ACN_BCR 和使用视觉对象标签的 ACN_BCL 模型,融入对抗学习[30]但分别使用区域特征、视觉对象特征、视觉对象标签的 ACN_GAN_BCR、ACN_GAN_BCL、ACN_GAN_BCO 等模型。实验结果如表 2 所示,表中 PER、LOC、ORG、MISC 分别代表数据集中的人名、地名、组织名和杂项等 4 类实体。

相比使用 Glove 的文本表示,使用 BERT 使文

表2 基于BiLSTM的MNER方法对比分析

Table 2 Comparative analysis of the MNER method based on BiLSTM %

模型	Twitter-2015						Twitter-2017					
	Single Type				Overall		Single Type				Overall	
	*F*1(PER)	*F*1(LOC)	*F*1(ORG)	*F*1(MISC)	*R*	*F*1	*F*1(PER)	*F*1(LOC)	*F*1(ORG)	*F*1(MISC)	*R*	*F*1
ACN_GCR[29]	82.00	79.00	53.10	34.00	68.70	70.70	—	—	—	—	—	—
VAM_GCR[28]	—	—	—	—	—	—	—	—	—	—	79.90	80.80
ACN_BR[36]	—	—	—	—	73.69	72.57	—	—	—	—	86.70	85.70
MA_BCR	84.83	80.24	59.14	36.65	74.48	72.75	91.20	83.91	82.58	67.74	85.86	85.01
VAM_BCR	85.21	80.41	60.60	38.33	**74.96**	72.77	92.09	82.65	82.59	67.28	86.38	85.12
ACN_BCR	85.48	79.81	60.74	36.90	73.91	73.04	91.21	85.30	83.70	68.94	86.08	85.55
ACN_BCL	84.97	80.26	60.77	38.02	74.40	73.22	91.40	82.91	83.80	62.82	85.20	84.79
ACN_GAN_BCR	**85.82**	80.47	60.94	**39.34**	74.48	72.95	91.78	**86.71**	**84.28**	**69.30**	**87.49**	**86.21**
ACN_GAN_BCL	85.09	79.12	**62.10**	37.71	74.02	72.95	**92.58**	82.02	83.77	63.37	85.42	85.29
ACN_GAN_BCO	85.31	**80.81**	60.48	38.72	74.87	**73.27**	91.87	85.64	83.50	67.33	86.68	85.85

本表示具有更完备的实体语义表示,因为BERT具备语言模型的背景知识。如在表2中VAM_GCR[28]、ACN_GCR[29]与ACN_BCR、VAM_BCR的实验对比中,后两者明显取得显著的优势。在Twitter-2017数据集中,VAM_BCR方法较VAM_GCR[28]方法R、$F1$值分别高出6.48百分点、4.32百分点,ACN_BCR的5项指标均高于ACN_GCR[29]方法。

将字符表示和单词表示进行拼接,通过补全单词表示中缺失的语义,以增强单词表示,进而得到更加准确的预测标签。ACN_BCR与ACN_BR[36]相比,在Twitter-2015数据集中R和$F1$值分别高出0.22百分点和0.47百分点,结果表明,使用文本模态内多特征协同表达,可以解决现有的文本表征模型存在语义缺失的问题。

由前融合模型MA、VAM与后融合模型ACN_BCR的对比中可以发现,在2个数据集中,ACN_BCR的12个指标均高于MA方法,10个指标均高于VAM方法。这表明使用BiLSTM融合单词表示和字符表示,使得文本表示具有更高的实体语义,能得到更好的多模态表示。

使用对抗学习实现2个表征空间的统一是有效的。对比ACN_GAN方法和ACN方法,11个最先进的性能指标出现在ACN_GAN方法,2个数据集中最高的$F1$值分别为ACN_GAN_BCO和ACN_GAN_BCR方法。这是因为对抗学习能使得文本表示和区域视觉特征的语义分布相似,从而更准确地融合,得到更高质量的多模态表示。

3 基于Transformer的MNER方法

基于Transformer的MNER方法使用Transformer模型和CRF作为基础模块,并使用BERT编码文本以缩小图文特征实体语义之间的差距。为解决视觉偏差的问题,利用多任务协同学习引导图像和文本特征深度融合,本文根据任务结构,划分为Transformer单任务模型和Transformer多任务模型,并对经典方法进行实验验证和方法分析。

3.1 Transformer单任务模型

Transformer单任务模型使用BERT进一步缩小文本与图像特征的语义差距,其处理流程是获得各模态的输入表示后采用多模态融合技术重新编码所有的模态输入表示,以获得多模态表示,最后通过CRF模型得到最终标签,具体框架如图4所示。Transformer单任务模型的核心是多模态融合技术,本小节将介绍所涉及的3种多模态融合技术路线。

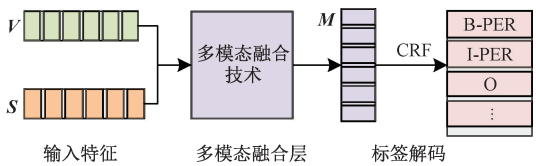

图4 Transformer单任务模型

Figure 4 Transformer single-task model

(1)图文联合编码。如Asgari-chenaghlu等[55]调用Transformer联合编码文本$\boldsymbol{S}$和图像分类标签$\boldsymbol{V}^{\mathrm{l}}$,将输出特征的文本部分作为多模态表示$\boldsymbol{M}$,可表示为

$$[\boldsymbol{M},\boldsymbol{V}'] = \mathrm{BERT}([\boldsymbol{S};\boldsymbol{V}^{\mathrm{l}}])。 \tag{25}$$

(2)感知表示融合技术。如Zhang等[31]将文本单词特征$\boldsymbol{X}^{\mathrm{b}} = \mathrm{BERT}(\boldsymbol{S})$和视觉对象$\boldsymbol{V}^{\mathrm{r}}$视作节点,分别使用模态内边连接模态内特征,使用模态间边连接模态间特征,构成无向图$\boldsymbol{G}$。然后堆叠n个基于图的跨模态注意力模型对$\boldsymbol{G}$进行编码,以实现特征融合,得到多模态表示$\boldsymbol{M}$。

$$\boldsymbol{G} = \text{Graph}(\boldsymbol{X}^{\text{h}}, \boldsymbol{V}^{\text{r}}); \tag{26}$$

$$[\boldsymbol{M}, \boldsymbol{V}''] = \text{cross-attention}(\boldsymbol{X}^{\text{h}}, \boldsymbol{V}^{\text{r}})。 \tag{27}$$

式中:Graph( ) 为将单词特征和视觉对象构建为无向图的函数;cross-attention( ) 为跨模态注意力机制,分别输出多模态表示 $\boldsymbol{M}$ 和多模态视觉表示 $\boldsymbol{V}''$。

钟维幸等[37]使用 ALBERT 分别对文本 $\boldsymbol{S}$ 和图像描述 $\boldsymbol{L}$ 进行编码,再使用由 3 个自注意力模型、4 个跨模态注意力模型组成的多模态融合模块来计算多模态表示 $\boldsymbol{M}$。

(3)多模态语义对齐技术。如 Xu 等[32]通过跨模态注意力模型先将文本特征和视觉对象对齐,得到匹配表示,再使用多模态注意力模型融合文本特征和视觉对象得到多模态表示。Liu 等[57]构建了多层次的对齐来捕获文本和图像之间由粗粒度到细粒度的交互,并通过计算文本和图像的相关性在不同语义层次上执行跨模态交互来增强文本表示,最终得到多模态表示。

3.2　Transformer 多任务模型

在单任务 MNER 模型的基础上,扩展了文本模态任务或其他辅助任务,以解决视觉偏差问题。Transformer 多任务模型的核心是多模态融合技术和多任务融合模块,框架如图 5 所示。

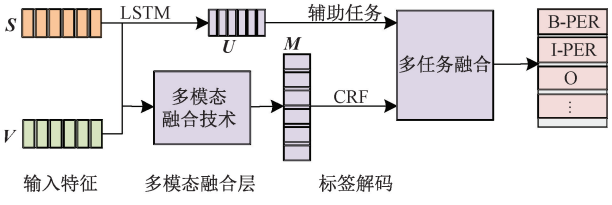

图 5　Transformer 多任务模型

Figure 5　Transformer multi-task model

多模态融合技术通常使用 Transformer 融合模态输入表示,得到多模态表示 $\boldsymbol{M}$。包含以下融合技术路线。

(1)感知表示融合技术。Yu 等[33]提出由 Transformer 模型对文本进行处理,多头跨模态注意力机制(multi-head cross-modal attention,MHCA)融合图文特征,得到多模态表示 $\boldsymbol{M}$,计算如下:

$$\boldsymbol{M} = \text{MHCA}(\text{Transformer}(\text{BERT}(\boldsymbol{S})), \boldsymbol{V}^{\text{r}})。 \tag{28}$$

式中:$\boldsymbol{S}$ 为输入文本;$\boldsymbol{V}^{\text{r}}$ 为区域视觉特征。

Liu 等[39]提出由 2 个 Transformer 模型分别对文本和视觉进行处理,然后使用多头跨模态注意力模型融合计算多模态表示 $\boldsymbol{M}$,计算如下:

$$\boldsymbol{M} = \text{MHCA}(\text{Transformer}(\text{BERT}(\boldsymbol{S})), \text{Transformer}(\boldsymbol{V}^{\text{r}}))。 \tag{29}$$

Zhang 等[58]基于 BERT 文本 token 和 SwinT 视觉 token 构造了多模态图,使用对比学习实现文本节点和视觉节点之间的全局对齐和局部对齐,之后堆叠 n 个跨模态注意力模型进行特征融合,得到多模态表示。

(2)图文联合编码。如 Wang 等[34]提出将图像描述 $\boldsymbol{V}^{\text{cap}}$、视觉对象 $\boldsymbol{V}^{\text{l}}$ 和文本 $\boldsymbol{S}$ 进行拼接,使用 BERT 进行编码得到多模态文本表示 $\boldsymbol{M}$,计算如下:

$$[\boldsymbol{M}, \boldsymbol{V}^{\text{lcap}}] = \text{BERT}([\boldsymbol{S}; \boldsymbol{V}^{\text{l}}; \boldsymbol{V}^{\text{cap}}])。 \tag{30}$$

Chen 等[41]提取层次视觉特征 $\boldsymbol{V}^{\text{g}}$,并将每层视觉特征和视觉对象特征进行拼接后,将其转换为视觉前缀 $\boldsymbol{V}^{\text{k}} = (\boldsymbol{V}_1^{\text{k}}, \boldsymbol{V}_2^{\text{k}}, \cdots, \boldsymbol{V}_{12}^{\text{k}})$,$\boldsymbol{V}^{\text{v}} = (\boldsymbol{V}_1^{\text{v}}, \boldsymbol{V}_2^{\text{v}}, \cdots, \boldsymbol{V}_{12}^{\text{v}})$,得到多模态表示 $\boldsymbol{M} = \boldsymbol{T}_{13}$,计算如下:

$$\boldsymbol{T}_{i+1} = \text{Transformerlayer}_i(\boldsymbol{T}_i, [\boldsymbol{V}_i^{\text{k}}, \boldsymbol{V}_i^{\text{v}}])。 \tag{31}$$

式中:$i = 1, 2, \cdots, 12$ 为 Transformer 编码层编号;$\boldsymbol{T}_i$ 为第 i 层的编码输出,$\boldsymbol{T}_1$ 为输入的文本编码。Transformer 编码层的自注意力机制(self-Attention,SA)计算原理表示为

$$\text{SA}(\boldsymbol{T}_i, \boldsymbol{V}_i^{\text{k}}, \boldsymbol{V}_i^{\text{v}}) = \text{softmax}\left(\frac{\boldsymbol{Q}_i \cdot [\boldsymbol{V}_i^{\text{k}}; \boldsymbol{K}_i]}{\sqrt{\boldsymbol{d}}}\right) \cdot [\boldsymbol{V}_i^{\text{v}}; \boldsymbol{V}_i]。 \tag{32}$$

式中:$\boldsymbol{Q}_i$、$\boldsymbol{K}_i$、$\boldsymbol{V}_i$ 均为 $\boldsymbol{T}_i$ 的投影向量。

多任务融合层通过联合优化模型参数,进而提升实体识别性能,主要包含多模态命名实体识别主任务结合文本模态任务或辅助任务的结构。多模态命名实体识别主任务是基于多模态表示的命名实体识别任务。文本模态任务是基于文本模态的解码任务,如 Yu 等[33]利用基于文本的实体跨度检测辅助预测;Wang 等[34]对齐多模态视图和文本视图的输出分布预测;Liu 等[39]通过不确定性判断使用文本候选标签或者多模态候选标签。

辅助任务能解决多模态表示过度融合视觉特征导致的偏差问题。如李晓腾等[40]所使用的对比学习、实体聚类辅助任务、边界检测任务;Chen 等[41]的关系抽取任务能优化多模态表示;Zhang 等[58]的多重对比学习任务能学习文本和图像表示的全局和局部一致性,从而过滤了语义不匹配或不相关的图文特征;Xu 等[59]提出数据鉴别器任务,将数据分给文本模态命名实体识别任务或多模态命名实体识别任务,获得最优的预测序列;Zhang 等[60]提出的硬样本挖掘策略,能优化文本和视觉特征对齐,缓减视觉对象的数量和类型所造成的偏差;Wang 等[61]提出的聚合命名实体分类任务和命名实体分割任务聚合视觉特征中的实体语义。

3.3　方法分析

在 Twitter-2015 和 Twitter-2017 2 个多模态数据

集上进行了实验,这2个数据集分别由Lu等[28]和Zhang等[29]提出,将每个数据集分割为训练数据集(Train)、验证数据集(Dev)、测试数据集(Test),分别统计数据集中的人名(PER)、地名(LOC)、组织名(ORG)、杂项(MISC)等4类实体的数量,统计数据如表3所示。

表3 2个Twitter的多模态数据集的统计数量

Table 3 Statistics of two multimodal Twitter datasets

实体类别	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
人名(PER)	2 217	552	1 816	2 943	626	621
地名(LOC)	2 091	522	1 697	731	173	178
组织名(ORG)	928	247	839	1 674	375	395
杂项(MISC)	940	225	726	701	150	157

通过评价指标R和$F1$值对基于Transformer的MNER模型的有效性进行对比分析。选取Transformer单任务模型中MSB[55]、UMGF[31]和MAF[32]模型,Transformer多任务模型中UMT[33]、ITA[34]和HvpNET[41]模型进行复现。在实验复现过程中,为了在同样的实验环境中运行,HvpNET批次大小降低为8,性能有所下降。具体实验结果如表4所示。

如表2和表4所示,在2个数据集中基于Transformer的方法整体优于基于BiLSTM的方法。这是因为Transformer的MNER方法使用BERT改进了文本表示,和Transformer融合模块具有同步挖掘上下文信息和融合多模态特征的能力,而基于BiLSTM的MNER方法是分阶段实现这2个功能的。如在Twitter-2015数据集中,基于BiLSTM方法的$F1$值最高值要低于基于Transformer方法的最低值,在Twitter-2017数据集中,相较VAM[28]、ACN[29]及ACN[36],基于Transformer方法的性能更好。

对ACN[36]进行改进,即通过融合字符表示或增加对抗学习任务,进一步补全文本语义以优化多模态表征后,ACN-GAN方法和MSB方法的性能相近。这表明使用BERT来表示文本语义仍然有缺失,而通过辅助任务可以促进多模态特征间的语义聚合,从而获得更好的多模态特征。

Transformer多任务模型通过任务间的共享学习和损失均衡,有效增强多模态表示通用性,同时也避免了多模态表示陷入局部最优值。由表4可以看出,在2个数据集上,Transformer多任务模型的各项指标均高于单任务模型,验证了多任务协同模型在多模态命名实体识别领域的优势。

视觉模态多特征协同对MNER具有重要作用。HvpNET协同使用层次视觉特征和对象级视觉特征作为文本的前缀特征,而UMT、UMGF、MAF仅使用区域视觉特征。在表4中可以发现,HvpNET方法的整体指标显著高于这3个方法。此外,使用视觉对象标签和图像标题的ITA方法,相比仅使用视觉对象标签的MSB方法,在2个数据集上的$F1$值分别高出2.13百分点、0.16百分点。这是因为2种视觉特征协同表示可以得到更全面的视觉语义。

进一步可以发现,相较于对象级视觉特征,使用自然语言对图像进行描述,图文语义能够更充分融合,如表4中,MSB的模型参数远小于UMT,却获得更好的性能。

3.4 模型的复杂度分析

模型参数量、单轮训练时间及单轮验证时间是衡量模型可用性的重要指标,对各模型进行统计,如表5所示。可以发现基于Transformer的MNER模型的参数量高于基于BiLSTM的MNER模型,综合$F1$值来看,模型参数量不是决定模型$F1$值的核心因素。Transformer多任务模型的训练时间与单任务模型的训练时间相当,但Transformer多任务模型的性能有较大提升。

表4 基于Transformer的MNER方法对比分析

Table 4 Comparative analysis of the MNER method based on Transformer %

分类	模型	Twitter-2015						Twitter-2017					
		Single Type				Overall		Single Type				Overall	
		$F1$(PER)	$F1$(LOC)	$F1$(ORG)	$F1$(MISC)	R	$F1$	$F1$(PER)	$F1$(LOC)	$F1$(ORG)	$F1$(MISC)	R	$F1$
Transformer单任务模型	MSB[55]	85.32	80.86	61.11	36.12	74.33	73.39	91.94	83.71	66.45	84.17	84.90	85.69
	UMGF[31]	84.68	80.49	58.5	40.16	73.44	72.93	90.47	81.56	66.44	82.04	83.73	84.16
	MAF[32]	84.55	80.38	61.6	40.99	74.86	73.09	90.76	85.63	64.24	84.68	84.60	85.33
Transformer多任务模型	ITA[34]*	—	—	—	—	—	75.52	—	—	—	—	—	85.85
	UMT[33]	85.2	80.73	61.63	42.22	75.11	73.47	90.88	85.88	63.87	83.19	84.97	84.88
	HvpNET[41]	85.74	81.78	61.92	40.81	75.65	74.33	92.04	84.42	66.88	85.13	85.79	86.14

注:*表示数据来自原文献。

表 5　不同模型的参数量、训练时间和验证时间对比

Table 5　Compare the number of parameters, training time and validation time of different models

模型	参数量/MB	训练时间/s	验证时间/s
MA_BCR	110.73	47.90	3.74
VAM_BCR	110.81	51.15	5.44
ACN_BCR	121.79	72.26	6.52
ACN_BCL	121.40	75.57	6.37
ACN_GAN_BCR	122.97	74.67	6.08
ACN_GAN_BCL	120.22	74.11	5.56
ACN_GAN_BCO	122.18	69.52	5.24
MA_BCR	110.73	47.90	3.74
VAM_BCR	110.81	51.15	5.44
MSB[55]	122.97	45.80	3.31
UMGF[31]	191.32	314.42	18.73
MAF[32]	136.09	103.39	6.37
ITA[34]	122.97	65.40	4.69
UMT[33]	148.10	156.73	8.59
HvpNET[41]	143.34	70.36	9.34

4　结束语

本文先对 MNER 任务的定义、难点及方法进行了简要介绍,然后总结了 MNER 方法框架,分别介绍框架中各部分的常用技术及其优缺点。接着对近年来 MNER 的方法进行梳理和分类,将其总结为 2 类方法和 4 种模型结构。为了评估基于 BiLSTM 的 MNER 方法,将其总结为前融合模型和后融合模型结构,在 Twitter-2015、Twitter-2017 数据集对 2 种模型结构中 7 种方法进行实验,分析如下:前融合模型是最早的 MNER 模型结构,该类模型在命名实体识别模型中添加视觉模态,并以简单、直接的方式实现多模态融合,其性能优于命名实体识别模型。后融合模型是前融合模型的改进,它初步解决了文本语义和图像语义不匹配的问题。笔者在后融合模型 ACN 的基础上进行拓展,解决现有文本表示方法中存在语义缺失问题,使用多特征协同表达,补全文本语义,性能进一步提升。

为了评估 Transformer 的 MNER 方法,将其总结为 Transformer 单任务模型、Transformer 多任务模型,在 Twitter-2015、Twitter-2017 数据集对 Transformer 单任务模型、Transformer 多任务模型中 6 种典型方法进行实验,分析如下:单任务模型使用 BERT 作为文本表示,利用 Transformer 实现多模态特征的深度融合,但存在视觉偏差问题,为此,通过利用文本表示或优化多模态表示的方法,将单任务模型扩展为多任务模型,其中包含 2 种多任务结构,即联合命名实体识别任务解决视觉偏差问题或聚合辅助任务增强多模态表示的通用性。

5　展望

本文对 4 类模型进行分析后,从以下 3 个方面指出了 MNER 未来的发展方向。

(1)多特征协同表达的重要性。模态内多特征协同表达能解决特征语义的问题,从而获得更加准确和全面的模态信息描述。

(2)多模态表征空间统一的重要性。当多模态特征空间统一,能解决融合特征时实体语义不匹配的问题。其中可以采用调用 Transformer 层对多模态表示重新编码,实现表征空间的统一和使用辅助任务优化特征投影,在投影空间中实现语义对齐。

(3)多任务学习的重要性。多任务模型与命名实体识别任务的结合是必要的,具体可以尝试以下几种研究思路:第一,使用多任务优化模态特征以利于编码、融合或对齐;第二,通过多任务协同学习通用的多模态表征,进而提升 MNER 性能;第三,结合迁移学习解决中文数据标注困难的问题,将多模态命名实体模型引入中文文本命名实体识别研究。

参考文献:

[1] GRISHMAN R, SUNDHEIM B. Message understanding conference-6: a brief history[C]//Proceedings of the 16th conference on Computational linguistics. Stroudsburg: ACL, 1996: 466-471.

[2] 余俊, 张学清. 音乐命名实体识别方法[J]. 计算机应用, 2010, 30(11): 2928-2931, 2948.
SHE J, ZHANG X Q. Musical named entity recognition method[J]. Journal of Computer Applications, 2010, 30(11): 2928-2931, 2948.

[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.
PAN Z G. Research on the recognition of Chinese named entity based on rules and statistics[J]. Information Science, 2012, 30(5): 708-712, 786.

[4] ZHOU G D, SU J. Named entity recognition using an HMM-based chunk tagger[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. New York: ACM, 2002: 473-480.

[5] 梁立荣, 李长伟, 沈晔, 等. 基于层叠条件随机场模型的电子病历文本信息抽取[J]. 计算机应用与软件, 2019, 36(10): 47-54, 112.
LIANG L R, LI C W, SHEN Y, et al. Text information extraction for electronic medical record based on cascaded conditional random field model[J]. Computer Applications and Software, 2019, 36(10): 47-54, 112.

[6] KONG J, ZHANG L X, JIANG M, et al. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2021, 116: 103737.

[7] 罗凌, 杨志豪, 宋雅文, 等. 基于笔画ELMo和多任务学习的中文电子病历命名实体识别研究[J]. 计算机学报, 2020, 43(10): 1943-1957.
LUO L, YANG Z H, SONG Y W, et al. Chinese clinical named entity recognition based on stroke ELMo and multi-task learning[J]. Chinese Journal of Computers, 2020, 43(10): 1943-1957.

[8] 杨飘, 董文永. 基于BERT嵌入的中文命名实体识别方法[J]. 计算机工程, 2020, 46(4): 40-45, 52.
YANG P, DONG W Y. Chinese named entity recognition method based on BERT embedding[J]. Computer Engineering, 2020, 46(4): 40-45, 52.

[9] 郭军成, 万刚, 胡欣杰, 等. 基于BERT的中文简历命名实体识别[J]. 计算机应用, 2021, 41(增刊1): 15-19.
GUO J C, WAN G, HU X J, et al. Chinese resume named entity recognition based on BERT[J]. Journal of Computer Applications, 2021, 41(S1): 15-19.

[10] 李博, 康晓东, 张华丽, 等. 采用Transformer-CRF的中文电子病历命名实体识别[J]. 计算机工程与应用, 2020, 56(5): 153-159.
LI B, KANG X D, ZHANG H L, et al. Named entity recognition in Chinese electronic medical records using transformer-CRF[J]. Computer Engineering and Applications, 2020, 56(5): 153-159.

[11] CETOLI A, BRAGAGLIA S, O'HARNEY A D, et al. Graph convolutional networks for named entity recognition[C]//Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories. Stroudsburg: ACL, 2018:37-45.

[12] TANG Z, WAN B Y, YANG L. Word-character graph convolution network for Chinese named entity recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1520-1532.

[13] SUI Y, BU F Y, HU Y T, et al. Trigger-GNN: a trigger-based graph neural network for nested named entity recognition[C]//2022 International Joint Conference on Neural Networks (IJCNN). Piscataway:IEEE, 2022: 1-8.

[14] 刘威,马磊,李凯,等.基于多粒度字形增强的中文医学命名实体识别[J].计算机工程,2024,50(2):337-344.
LIU W, MA L, LI K, et al. Chinese medical named entity recognition based on multi-granularity glyph enhancement[J]. Computer Engineering, 2024, 50(2): 337-344.

[15] 赵珍珍,董彦如,刘静等.融合词信息和图注意力的医学命名实体识别[J/OL].计算机工程与应用,2023:1-11(2023-06-14)[2023-09-27]. https://kns.cnki.net/kcms2/detail/11.2127.TP.20230613.1328.010.html.
ZHAO Z Z, DONG Y R, LIU J, et al. Medical named entity recognition incorporating word lnformation and graph attention [J/OL]. Computer Engineering and Applications, 2023, 1-11(2023-06-14)[2023-09-27]. https://kns.cnki.net/kcms2/detail/11.2127.TP.20230613.1328.010.html.

[16] 陈曙东, 罗超, 欧阳小叶, 等. 基于动态词典匹配的语义增强中文命名实体识别算法[J]. 无线电工程, 2021, 51(7): 519-525.
CHEN S D, LUO C, OUYANG X Y, et al. A semantic-enhanced Chinese named entity recognition algorithm based on dynamic dictionary matching[J]. Radio Engineering, 2021, 51(7): 519-525.

[17] 胡新棒, 于溆乔, 李邵梅, 等. 基于知识增强的中文命名实体识别[J]. 计算机工程, 2021, 47(11): 84-92.
HU X B, YU X Q, LI S M, et al. Chinese named entity recognition based on knowledge enhancement[J]. Computer Engineering, 2021, 47(11): 84-92.

[18] 耿志超, 颜航, 邱锡鹏, 等. 基于不确定片段的检索增强命名实体识别框架[J]. 中文信息学报, 2023, 37(7): 71-81.
GENG Z C, YAN H, QIU X P, et al. The uncertainty-based retrieval framework for Chinese NER[J]. Journal of Chinese Information Processing, 2023, 37(7): 71-81.

[19] 廖梦,贾真,李天瑞.基于标签信息融合与多任务学习的中文命名实体识别[J/OL].计算机科学, 2023:1-11(2023-09-26)[2023-09-27]. https://link.cnki.net/urlid/50.1075.TP.20230925.2014.235.
LIAO M, JIA Z, LI T R. Chinese named entity recognition based on label information fusion and multi-task learning[J]. Computer Science, 2023:1-11(2023-09-26)[2023-09-27]. https://link.cnki.net/urlid/50.1075.TP.20230925.2014.235.

[20] 王蓬辉, 李明正, 李思. 基于数据增强的中文医疗命名实体识别[J]. 北京邮电大学学报, 2020, 43(5): 84-90.
WANG P H, LI M Z, LI S. Data augmentation for Chinese clinical named entity recognition[J]. Journal of Beijing University of Posts and Telecommunications, 2020, 43(5): 84-90.

[21] 余传明, 林虹君, 张贞港. 基于多任务深度学习的实体和事件联合抽取模型[J]. 数据分析与知识发现, 2022, 6(增刊1): 117-128.
YU C M, LIN H J, ZHANG Z G. Joint extraction model for entities and events with multi-task deep learning[J].

Data Analysis and Knowledge Discovery, 2022, 6(S1): 117-128.

[22] 武国亮, 徐继宁. 基于命名实体识别任务反馈增强的中文突发事件抽取方法[J]. 计算机应用, 2021, 41(7): 1891-1896.
WU G L, XU J N. Chinese emergency event extraction method based on named entity recognition task feedback enhancement[J]. Journal of Computer Applications, 2021, 41(7): 1891-1896.

[23] ARSHAD O, GALLO I, NAWAZ S, et al. Aiding intra-text representations with visual context for multimodal named entity recognition[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). Piscataway:IEEE, 2019: 337-342.

[24] ESTEVES D, PERES R, LEHMANN J, et al. Named entity recognition in twitter using images and text[C]//International Conference on Web Engineering. Cham: Springer, 2018: 191-199.

[25] CHEN D W, LI Z X, GU B B, et al. Multimodal named entity recognition with image attributes and image knowledge[C]//International Conference on Database Systems for Advanced Applications. Cham: Springer, 2021: 186-201.

[26] 范涛, 王昊, 陈玥彤. 基于深度迁移学习的地方志多模态命名实体识别研究[J]. 情报学报, 2022, 41(4): 412-423.
FAN T, WANG H, CHEN Y T. Research on multimodal named entity recognition of local history based on deep transfer learning[J]. Journal of the China Society for Scientific and Technical Information, 2022, 41(4): 412-423.

[27] 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. New Orleans: NAACL, 2018:852-860.

[28] 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: ACL, 2018: 1990-1999.

[29] 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. Palo Alto: AAAI,2018:5674-5681.

[30] 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, 2020, 23: 2520-2532.

[31] ZHANG D, WEI S Z, LI S S, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2021: 14347-14355.

[32] 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.

[33] 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: ACL, 2020: 3342-3352.

[34] 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: ACL, 2022: 3176-3189 .

[35] 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.

[36] CHEN S G, AGUILAR G, NEVES L, et al. Can images help recognize entities? a study of the role of images for Multimodal NER[C]//Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021). Stroudsburg: ACL, 2021: 87-96.

[37] 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 multi-modal semantic interaction[J]. Guangxi Sciences, 2022, 29(4): 681-690.

[38] TIAN Y, SUN X, YU H F, et al. Hierarchical self-adaptation network for multimodal named entity recognition in social media[J]. Neurocomputing, 2021, 439: 12-21.

[39] 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.

[40] 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.

[41] 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 ACL: NAACL 2022. Stroudsburg: ACL, 2022: 1607-1618.

[42] SUI D B, TIAN Z K, CHEN Y B, et al. A large-scale Chinese multimodal NER dataset with speech clues[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: ACL, 2021: 2807-2818.

[43] LIU Y, HUANG S B, LI R S, et al. USAF: multimodal Chinese named entity recognition using synthesized acoustic features[J]. Information Processing & Management, 2023, 60(3): 103290.

[44] 冯皓楠, 何智勇, 马良荔. 基于图文注意力融合的主题标签推荐[J]. 郑州大学学报(工学版), 2022, 43(6): 30-35.
FENG H N, HE Z Y, MA L L. Multimodal hashtag recommendation based on image and text attention fusion [J]. Journal of Zhengzhou University (Engineering Science), 2022, 43(6): 30-35.

[45] 郑建兴, 郭彤彤, 申利华, 等. 基于评论文本情感注意力的推荐方法研究[J]. 郑州大学学报(工学版), 2022, 43(2): 44-50, 57.
ZHENG J X, GUO T T, SHEN L H, et al. Research on recommendation method based on sentimental attention of review text[J]. Journal of Zhengzhou University (Engineering Science), 2022, 43(2): 44-50, 57.

[46] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2023-01-16)[2023-06-18]. https://arxiv.org/abs/1301.3781.pdf.

[47] GOLDBERG Y, LEVY O. Word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method[EB/OL]. (2014-02-15)[2023-06-18]. https://arxiv.org/abs/1402.3722.pdf.

[48] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2014: 1532-1543.

[49] ATHIWARATKUN B, WILSON A, ANANDKUMAR A. Probabilistic FastText for multi-sense word embeddings [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2018: 1-11.

[50] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2018: 2227-2237.

[51] ZHONG Q, TANG Y. An attention-based BILSTM-CRF for Chinese named entity recognition[C]//2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA). Piscataway:IEEE, 2020: 550-555.

[52] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE, 2016: 770-778.

[53] 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.

[54] 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.

[55] 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.

[56] LIU Y G, ZHOU Y M, WEN S T, et al. A strategy on selecting performance metrics for classifier evaluation[J]. International Journal of Mobile Computing and Multimedia Communications, 2014, 6(4): 20-35.

[57] LIU P P, LI H, REN Y M, et al. A novel framework for multimodal named entity recognition with multi-level alignments[EB/OL]. (2023-05-15)[2023-06-18]. https://doi.org/10.48550/arxiv.2305.08372.

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

[59] XU B, HUANG S, DU M, et al. Different data, different modalities! reinforced data splitting for effective multimo-

dal information extraction from social media posts[C]//Proceedings of the 29th International Conference on Computational Linguistics. Stroudsburg: ACL, 2022:1855-1864.

[60] 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.

[61] 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.

Research Progress of Multimodal Named Entity Recognition

WANG Hairong [1,2], XU Xi[1], WANG Tong[1], JING Boxiang[1]

(1. College of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China; 2. The Key Laboratory of Images & Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China)

Abstract: In order to solve the problems in studies of multimodal named entity recognition, such as the lack of text feature semantics, the lack of visual feature semantics, and the difficulty of graphic feature fusion, a series of multimodal named entity recognition methods were proposed. Firstly, the overall framework of multi modal named entity recognition methods and common technologies in each part were examined, and classified into BilSTM-based MNER method and Transformer based MNER method. Furthermore, according to the model structure, it was further divided into four model structures, including pre-fusion model, post-fusion model, Transformer single-task model and Transformer multi-task model. Then, experiments were carried out on two data sets of Twitter-2015 and Twitter-2017 for these two types of methods respectively. The experimental results showed that multi-feature cooperative representation could enhance the semantics of each modal feature. In addition, multi-task learning could promote modal feature fusion or result fusion, so as to improve the accuracy of MNER. Finally, in the future research of MNER, it was suggested to focus on enhancing modal semantics through multi-feature cooperative representation, and promoting model feature fusion or result fusion by multi-task learning.

Keywords: multimodal named entity recognition; Transformer; BiLSTM; multimode fusion; multitasking learning