

基于深度迁移学习的 地方志多模态命名实体识别研究

范 涛, 王 昊, 陈玥彤

(南京大学信息管理学院, 南京 210023)

摘 要 地方志作为中华文化的组成部分, 是建设文化强国的重要一环, 对其进行挖掘研究具有重要意义; 同时, 有效识别实体对地方志知识组织和知识图谱构建有着重要影响。当前地方志命名实体识别研究主要基于文本, 缺乏文本对应的图片, 而图片中的内容能够为识别文本中的实体提供额外的信息, 从而提升模型识别实体的性能, 并且实体识别还面临着已标注语料匮乏的问题。基于此, 本文提出了利用深度迁移学习方法, 结合地方志中的文本和图片进行多模态命名实体识别。首先, 基于人民日报语料库和中文推特多模态数据集, 分别预训练结合了自注意力机制的 BiLSTM-attention-CRF 模型和自适应联合注意力模型, 利用基于神经网络的深度迁移学习方法将权重迁移至地方志多模态命名识别模型中, 使模型获得提取文本和图片语义特征的能力; 然后, 结合过滤门对多模态融合特征去噪; 最后, 将融合后的多模态特征输入 CRF (conditional random fields) 层进行解码。本文将提出的模型在地方志多模态数据中进行了实证研究, 并同相关基线模型作对比, 实验结果表明, 本文所提出的模型具有一定优势。

关键词 深度迁移学习; 多模态命名实体识别; 地方志

Research on Multimodal Named Entity Recognition of Local History Based on Deep Transfer Learning

Fan Tao, Wang Hao and Chen Yuetong

(School of Information Management, Nanjing University, Nanjing 210023)

Abstract: Local history, as a part of Chinese culture, is important in constructing culture in China, making it meaningful to examine local history. Recognizing entities has a great impact on knowledge organization and the construction of knowledge graphs. Researches on the Named Entity Recognition (NER) of local history are based on texts but lacking the combination of texts and images. Images can provide extra information to recognize the entity, increasing the performance of the NER model. Additionally, the NER model still lacks annotated corpus. Thus, in this study we proposed a multimodal NER model based on deep transfer learning, combining texts and images. First, we pretrained BiLSTM-attention-CRF and adaptive co-attention model based on Renmin Daily corpus and Chinese Twitter multimodal data. Then, we used a method based on deep neural network to transfer the weights in pretrained models to the proposed model, making the proposed multimodal NER model able to capture the semantics features in texts and images. A filter gate was applied to filter the information from the multimodal features. Lastly, a CRF layer was adopted to decode the fused multimodal features, outputting the labels. The proposed multimodal NER model was evaluated on the local history multimodal dataset and compared

收稿日期: 2021-02-18; 修回日期: 2021-05-07

基金项目: 国家自然科学基金面上项目“关联数据驱动下我国非遗文本的语义解析与人文计算研究”(72074108); 南京大学文科青年跨学科团队专项“面向人文计算的方志文本的语义分析和知识图谱研究”(010814370113)。

作者简介: 范涛, 男, 1995 年生, 博士研究生, 主要研究领域为自然语言处理; 王昊, 男, 1981 年生, 博士, 教授, 博士生导师, 主要研究领域为自然语言处理, E-mail: ywhaowang@nju.edu.cn; 陈玥彤, 女, 1998 年生, 硕士研究生, 主要研究领域为文献计量。

with baseline methods. Experimental results showed the superiority of the proposed model.

Key words: deep transfer learning; multimodal named entity recognition; local history

1 引言

党的十九届五中全会通过的《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》中明确提出了到2035年建成文化强国的远景目标，并强调在“十四五”时期推进社会主义文化强国建设，这标志着我国文化强国建设进入了一个新的历史阶段^[1]。作为中华文化的载体和组成部分，地方志是建设文化强国的重要一环，对其进行挖掘和研究，有利于传播中华文化和增强文化自信^[2]。

命名实体识别作为文本挖掘中的一项基础任务，旨在识别文本中的专有词，如人名、地名、时间、组织等，其对后续的文本知识组织和知识图谱的构建都具有重要影响^[3]。目前，已有学者利用相关研究方法对地方志等文化资源进行了实体抽取。例如，李娜^[4]以《方志物产》山西分卷作为语料，基于条件随机场模型实现了对物产别名实体的自动抽取。黄水清等^[5]将部分人工标注的先秦古汉语语料库作为条件随机场的训练数据，利用训练生成的最优模型，对语料库中的地名实体进行自动识别。从上述工作可以看出：①当前对于地方志等文化资源命名实体识别任务的研究对象均基于文本，缺乏对多模态内容（即文本结合图片）的探究；②自动识别文本实体的模型依赖于大规模人工标注的语料，需要耗费大量的人力资源和时间。然而，随着地方志数字化进程的加快，地方志数据库提供的内容并不仅局限于文本这一单模态内容，与文本相关的图片资源同样以结构化的方式呈现，这为地方志多模态内容的研究提供了契机。在文本命名实体识别任务中，当实体边界模糊时，仅依靠上下文难以辨别其实体类型。例如，在图1中，倘若仅考虑文本，难以确定句子中所包含实体的边界，“江大桥”可以被视作人名，而“长江大桥”又可以被视作地名，但是当结合文本对应的图片时，则可以确定文本中提及的实体为“长江大桥”，从而准确地识别出实体。当面向某一具体领域展开实体识别研究时，通常会面临标注语料匮乏的问题。常用的解决方法是利用人工去标注数据集，但是会耗费大量的人力、物力，同时，在面向新领域时，还需标注新的语料，并不能较好地解决面向特定领域的实体

识别问题。然而，通过深度迁移学习方法，利用深度神经网络预学习相关领域知识后，再对目标语料进行实体抽取，则可以有效避免对训练语料的标注。目前，已有学者利用基于深度迁移学习的方法抽取文本中的实体，应用公开数据集训练模型，结合微调的方法提升实体抽取模型的性能^[6-7]。但是，目前的相关研究多集中于文本，利用深度迁移方法对多模态内容进行命名实体识别鲜有探索。基于此，为了解决目标领域标注数据匮乏的问题以及提升实体识别性能，本文提出利用深度迁移学习并结合文本和图片内容展开地方志多模态命名实体识别的研究。



图1 南京市长江大桥

来源于《南京交通志》，海天出版社1994年出版。

多模态命名实体识别是一项新兴的任务，旨在利用多模态内容挖掘文本和图片中存在的相关语义关系，增强文本语义信息，提升模型识别实体的性能。该任务最早由Zhang等^[8]提出，其利用基于自适应多模态联合注意力机制（adaptive co-attention）的命名实体识别模型，对推特中网民所发布的包含多模态内容的帖子进行实体识别，并获得了最优结果；同时作者公开了文中所用的多模态数据集。目前，中文领域尚未有应用于多模态命名实体识别的公开数据集，因此，本文以文献[8]的数据集为基础，制作了用于深度迁移学习的平行语料。尽管图片内容能够在一定程度上提升命名实体识别任务的性能，但是文本中的语义信息依旧是实体抽取中的核心。基于此，本文提出基于深度迁移学习的多模态命名实体识别模型（multimodal named entity recognition model, MNERM）。该模型主要由四个部分组成，分别是BiLSTM-attention module（BAM）模

块、adaptive co-attention module (ACAM) 模块、过滤门及CRF (conditional random fields) 层。为使得BAM模块和ACAM模块分别获取预训练权重, 本文分别引入了面向人民日报语料库的BiLSTM-attention-CRF (BAC) 模型和面向中文平行推特多模态语料库的adaptive co-attention CRF (ACAC) 模型, BAM模块和ACAM模块同样也是BAC模型和ACAC模型的组成部分。通过在对对应语料库预训练模型, 将权重参数分别迁移至BAM模块和ACAM模块, 使MNERM模型拥有提取多模态特征的能力。尽管应用多模态特征能够提升模型性能, 但依旧包含噪声, 本文提出利用过滤门对ACAM模块输出的多模态特征进行去噪, 再同BAM模块输出的文本特征进行融合, 最后以微调的方式将融合后的多模态特征输入至CRF层进行解码。

本文的主要贡献为: 从多模态视角出发, 提出结合地方志中的文本和图片进行命名实体的识别研究; 针对目标领域标注语料匮乏的问题, 提出利用深度迁移学习方法进行地方志多模态命名实体识别, 并构建了MNERM模型, 该模型能够充分获取不同模态的信息表示, 并能有效捕捉不同模态间的相关关系, 增强文本的特征表示能力。

本文将提出的模型在地方志多模态数据集中进行了实证研究, 并与相关基线模型进行对比。研究结果表明, 本文提出的模型具有一定的优越性。

2 相关研究

2.1 地方志命名实体识别研究

伴随着数字化进程的加快, 沉睡的人文资源逐步成为可计算的数据, 这为数字人文计算打下坚实的基础。而命名实体识别作为自然语言处理中的基础性任务, 其对文本的知识组织及实体间的关系抽取都有着重要的影响。为了探究古籍方志中的实体自动识别, 徐晨飞等^[9]采用BiLSTM-CRF、BERT等模型对物产别名、人物、产地及引书等实体进行识别, 实验结果表明, 采用基于深度学习的实体识别方法能够取得较好的效果。崔竞峰等^[10]基于深度学习方法, 构建BiLSTM-CRF模型对菊花古典诗词中的菊花花名、花色等实体进行识别, 并同CRF等基线模型作对比, 实验结果表明, 该文献提出的方法能够取得较好的效果。史书中的历史事件名是历史文本知识库的重要组成部分, 唐慧慧等^[11]提出以字作为最小语义单元, 利用CRF模型对魏晋南北朝史

书文中的历史事件名实体进行识别, 并取得良好效果。在人民日报语料库中, 殷章志等^[12]利用基于BiLSTM-CRF的序列标注模型抽取文本序列的中间特征, 并将其输入支持向量机中进行实体识别, 并取得一定的效果。石春丹等^[13]提出利用双向门控循环网络与CRF结合的模型对文本中人名、地名和机构名等实体进行识别, 该模型能够有效学习序列的时序信息, 并能捕捉长距离依赖。

从上述研究可以看出, 目前面向地方志等人文资源的命名实体识别研究多基于文本, 并利用基于BiLSTM-CRF架构的深度学习模型进行实体识别。与之不同的是, 本文在BAC模型中引入了自注意力机制, 其能够有效增强文本的特征表示, 减少序列信息中的噪声, 并获得实体识别性能上的提升。除此之外, 人文资源的数字化带来的并不止是单一的文本, 同时有着大量可获取的对应图片资源。已有研究表明, 图片的加入能在一定程度上增强和补充对应的文本语义信息^[14]。基于此, 本文提出结合地方志中的文本和图片, 进行命名实体识别研究。

2.2 多模态命名实体识别研究

用户在网络中产生内容的多模态化, 为多模态自然语言处理任务提供了丰富资源。多模态命名实体识别作为其中的一项任务, 已受到学界和工业界的广泛关注。在以文本为主要处理对象的命名实体任务中, 当实体存在多义性或实体边界难以区分时, 仅依靠上下文对实体类别做出准确判断存在一定困难。但是当文本有着与之对应的图片时, 通过观察图片内容出现的实体, 则能对歧义实体做出准确预测。

在多模态命名实体识别中, 文本和图片存在语义相关关系。在图片内容中, 与文本中提及实体存在相关关系的仅局限于图片的部分区域。因此, Zhang等^[8]提出基于自适应联合注意力机制 (adaptive co-attention) 的多模态命名实体模型, 利用词引导和图引导的注意力机制充分学习文本和图片的语义相关关系及模态交互, 应用门机制进行多模态特征融合和噪声过滤, 之后将多模态特征与经过BiLSTM编码后的文本特征再次拼接, 获得最终多模态融合特征, 并将其输入CRF层中进行解码, F1值达到70.69%。同样地, 为了充分学习图片中与文本实体相对应的语义特征, Yu等^[14]提出基于Transformer架构的多模态命名实体模型, 该模型主要由单模态特征表示、多模态Transformer及辅助实体边

界检测组成,通过这些构件,模型能够较好地学习文本和图片上下文敏感特征,并能够关注到聚合多模态信息时未被充分关注的实体。为充分理解图片中的视觉内容,Lu等^[15]提出基于视觉注意力机制的多模态命名实体模型,该模型能够自动忽略与文本内容无关的视觉信息并重点关注与文本内容最相关的视觉信息,其在多个数据集中取得较好结果。

上述研究主要通过挖掘图片与文本之间的相关语义关系及不同模态间的交互,并结合注意力机制,在公开英文数据集中取得一定性能。然而,在中文领域中,多模态命名实体识别任务尚未有研究涉及,并且缺乏相关的中文多模态命名实体识别语料。因此,本文探索将公开的英文多模态命名实体识别语料库制作成可学习的平行中文多模态命名识别语料库,并将词作为句子的划分粒度,利用深度迁移学习的方法对地方志多模态数据集进行实体识别研究。

2.3 深度迁移学习研究

深度迁移学习常用的方法包括基于实例的深度迁移学习(instance-based deep transfer learning)、基于映射的深度迁移学习(mapping-based deep transfer learning)、基于神经网络的深度迁移学习(network-based deep transfer learning)以及基于对抗的深度迁移学习(adversarial-based deep transfer learning)^[16]。其基本思想是利用在源域(source domain)训练的深度神经网络中的知识解决目标域(target domain)

中的问题。

目前,已有相关文献利用深度迁移学习方法进行命名实体识别研究。武惠等^[17]提出利用基于实例的深度迁移方法学习样本特征,构建BiLSTM-CRF模型对人民日报语料库中的实体进行识别,并取得一定效果。王瑞银等^[7]在源域中训练语言模型预测模型,将源域模型知识迁移至目标域模型中,从而对实体进行识别,其在法律文书数据集中性能良好。为了缓解可利用标注语料的不足, Lee等^[6]提出在大型源数据集中训练BiLSTM-CRF实体识别模型,结合微调的方法对目标域的实体进行识别,并取得了一定的效果。

为了有效获取文本的语义知识和文本结合图片的多模态知识,本文应用基于神经网络的深度迁移学习思想,提出在两个源域数据集中训练与目标模型对应部分有着相似结构的深度学习模型,然后将预训练模型中的权重迁移至目标模型的对应结构中,最后结合微调的方法对地方志多模态数据进行实体识别。

3 模型设计

为了提升地方志中模型识别实体的性能并探索解决目标领域标注语料匮乏问题,本文提出基于深度迁移学习的多模态命名实体模型MNERM,结构具体如图2所示,其分别由BAM模块、ACAM模块、过滤门及CRF层组成。本文首先分别在人民日

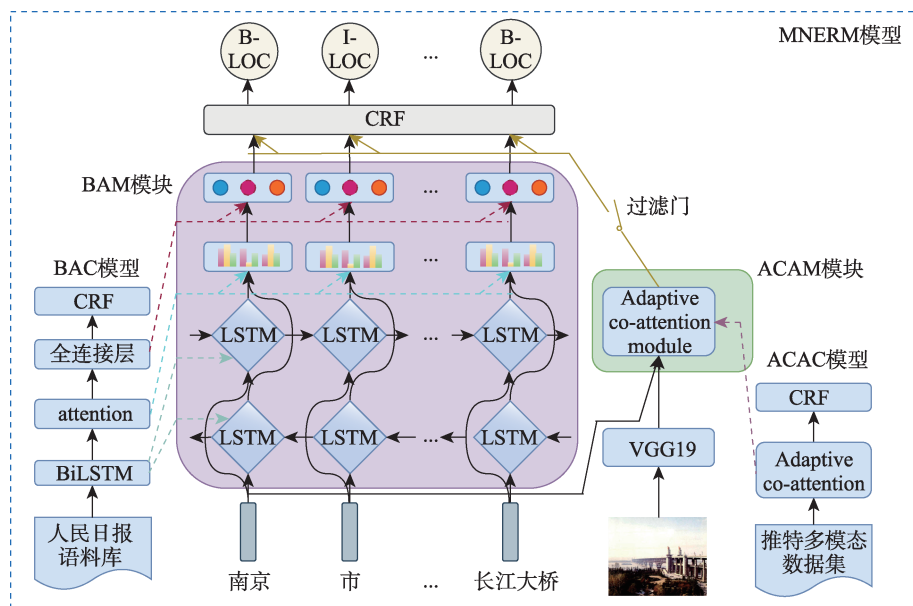


图2 基于深度迁移学习的多模态命名实体识别模型
虚线表示权重迁移。

报语料库和中文推特多模态数据集这两个源域预训练 BAC 模型和 ACAC 模型。然后, 利用基于神经网络的深度迁移学习方法, 将 BAC 模型和 ACAC 模型中的对应权重分别迁移至 BAM 模块和 ACAM 模块中, 使得 MNERM 具备抽取文本和图片的多模态特征能力。接着, 将文本特征和经过过滤门过滤的多模态特征进行中间层融合, 输入 CRF 层中进行解码生成标签, 并进行微调。下文将详述 MNERM 模型及建模方法。

3.1 特征提取

1) 文本特征提取

文本的特征表示对下游任务的表现有着重要影响。本文利用在百度百科大规模语料中预训练的中文词向量模型^[18], 对文本进行特征表示。MNERM 模型以 Skip-Gram 模型为基础, 并结合负采样技术进行优化, 其在中文类比推理任务中取得最优结果。本文利用 MNERM 模型分别对人民日报语料库、中文推特多模态语料及地方志多模态语料库中的句子进行文本表示。

2) 图片特征提取

以卷积神经网络 (convolutional neural network, CNN)^[19]为基础构建的模型, 如 VGG16、VGG19 等^[20], 在多个计算机视觉任务中均获得了最优结果。这一方面得益于 CNN 强大的特征学习建模能力, 另一方面则受益于大规模的图片训练集, 如 ImageNet^[21]。目前常用的图片提取方法是利用 ImageNet 数据集中预训练的 CNN 模型, 提取最后一层全连接层的输出作为图片的特征表示。但为了获取图片的空间特征表示, 本文遵循文献[8]中的方法, 以预训练于 ImageNet 数据集的 VGG19 模型中的最后一层池化层的输出作为图片的特征表示。本文利用 MNERM 模型分别提取中文推特多模态语料及地方志多模态语料中的图片特征。

3.2 BiLSTM-attention-CRF 模型

文本的语义信息是识别实体类别的核心, 已有研究表明, 将人民日报语料库 (1988) 作为迁移学习的学习语料, 并利用基于深度迁移学习的方法对其他语料库中的相同实体进行识别, 有着良好的效果^[17]。为了使 MNERM 模型中的 BAM 模块拥有先验知识, 本文设计了用于权重迁移的 BAC 模型。目前常用的命名实体模型多基于 BiLSTM-CRF 架构^[7-8], 与之不同的是, 本文引入了自注意力机制

(self-attention), 而利用自注意力机制能够有效增强文本的语义表示。BAC 模型主要由 BiLSTM 网络、自注意力层及 CRF 层。BAM 模块由 BAC 模型中的 BiLSTM 网络和自注意力层组成。BiLSTM 作为循环神经网络 (recurrent neural network, RNN) 的变体, 能够较好地学习句子中的上下文关系, 具有捕捉长距离依赖的能力, 并能够克服因序列长度过长所带来的梯度消失和梯度爆炸的问题。给定人民日报语料库中的句子 $S = \{s_1, s_2, \dots, s_i, \dots, s_n\}$, 进行特征表示后的句子为 $S' = \{s'_1, s'_2, \dots, s'_i, \dots, s'_n\}$, $s'_i \in \mathbb{R}^{d_s}$, 其中, n 表示句子长度, d_s 表示向量维度, 大小为 300。BiLSTM 获得的隐藏层状态 $h_i \in \mathbb{R}^d$ 由前向的 LSTM 输出 \vec{h}_i 和反向的 LSTM 输出 \overleftarrow{h}_i 拼接而成, d 表示隐藏层单元数, 具体公式为

$$\vec{h}_i = \text{LSTM}(s'_i, \vec{h}_{i-1}) \quad (1)$$

$$\overleftarrow{h}_i = \text{LSTM}(s'_i, \overleftarrow{h}_{i+1}) \quad (2)$$

$$h_i = [\vec{h}_i; \overleftarrow{h}_i] \quad (3)$$

注意力机制起源于人类视觉, 当人观察物体或阅读书本时, 会对其中的某一区域投入大量注意力, 获取富含价值的信息, 并抑制对其他区域的注意力投入。目前已有工作利用注意力机制进行自然语言处理任务, 如机器翻译、情感分析等; 而有关利用自注意力机制进行命名实体识别任务的研究相对较少。通过利用注意力机制, 能够确定在决定词的标签时, 有多少词的信息被利用, 从而提升模型性能。自注意力机制关注句子内部的特征相关性, 并能够减少对外部特征的依赖。在自注意力机制中, 句子中的每个语义单元同其他语义单元进行注意力权重计算, 可以有效捕捉词间的相互关系, 获取句子结构信息, 增强特征表示。自注意力机制本质上是输入 Query (Q) 到一系列键值对 (Key (K), Value (V)) 的映射函数, 对 BiLSTM 生成的句子表示 $H = \{h_i | h_i \in \mathbb{R}^d, i = 1, 2, \dots, n\}$, 应用自注意力机制获得的编码表示为 $E = \{e_i | e_i \in \mathbb{R}^d, i = 1, 2, \dots, n\}$, 具体公式为

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \times V \quad (4)$$

其中, Q 、 K 、 V 为隐藏层状态 h_i 的特征; Softmax 为归一化函数。将编码后的文本表示输入 CRF 层进行解码, 获得文本中词对应的预测标签 $Y = \{y_1, y_2, \dots, y_i, \dots, y_n\}$,

$$Y = \operatorname{argmax}_{y' \in y(E)} \frac{\prod_{i=1}^n \exp(W_{y_{i-1}, y_i}^T E + b_{y_{i-1}, y_i})}{\prod_{i=1}^n \exp(W_{y'_{i-1}, y'_i}^T E + b_{y'_{i-1}, y'_i})} \quad (5)$$

其中, W 、 b 表示全权重矩阵。本文利用经典的极大条件似然估计对 CRF 层进行训练, 具体公式为

$$L(W, b) = \sum_{(e_i, y_i)} \log p(y_i | e_i; W, b) \quad (6)$$

3.3 自适应联合注意力机制模型

鉴于当前尚未有中文多模态命名实体识别公开数据集, 仅有英文推特多模态命名实体识别公开数据集, 目前已有研究涉及利用英译汉平行语料来进行深度迁移学习, 并在公开数据集中取得了较好的性能^[22]。因此, 本文制作了推特多模态数据集的中文平行语料作为 ACAC 模型的训练语料, 将 ACAC 模型中自适应联合注意力网络的权重灌入 ACAM 模块中, 其主要由自适应联合注意力机制网络和 CRF 层组成。不同于自适应联合注意力机制结构^[8], 在 ACAC 模型中, 本文将 VGG-16 图片特征提取模型替换成性能更佳的 VGG-19^[23], 其余部分保持一致。

自适应联合注意力机制由词引导的注意力机制 (word-guided attention, WGA)、图引导的注意力机制 (image-guided attention, IGA) 和门机制组成。由图 1 可以看出, 图片中仅包含长江大桥的区域与文本中的“长江大桥”有关, 如果考虑图片中的全部区域, 那么会带来噪声和信息冗余。词引导的注意力机制核心思想是给序列中的一个词, 利用 Softmax 函数计算图片中的各个区域同该词的相关程度, 过滤掉与其不相关的区域和信息, 减少计算复杂度, 以达到最优结果。应用词引导的注意力机制, 则能让模型过滤掉噪声并找出与当前词最为相关的图片区域。给定文本序列 $X = \{x_1, x_2, \dots, x_i, \dots, x_n\}$, 利用 BiLSTM 编码后的输出表示为 $M = \{m_i | m_i \in \mathbb{R}^d, t = 1, 2, \dots, n\}$, 利用 VGG19 模型提取与文本相对应的图片特征为 $T = \{t_i | t_i \in \mathbb{R}^{512}, i = 1, 2, \dots, 49\}$, 其中特征图的数量为 49, 512 表示特征图的维度。应用词引导的注意力机制得到与词 m_i 相关的图片特征向量 \hat{t}_i ,

$$\hat{t}_i = \text{WGA}(\theta_w; m_i, t_i) \quad (7)$$

其中, θ_w 为词引导的注意力机制中的参数。利用 WGA 能够获得与词 m_i 相关的图片特征向量 \hat{t}_i , 但是并不知道序列中的哪个词与 m_i 相关。因此, 需要利用图引导的注意力机制去寻找与图片特征的最相关的词。图引导的注意力机制的核心思想是在给

定新的图片特征向量下, 计算序列中的词同图片特征向量的相关程度, 从而提升序列的特征表达能力。因此, 利用 IGA 可以计算出与图片特征表示 \hat{t}_i 相关的词 \hat{m}_i ,

$$\hat{m}_i = \text{IGA}(\theta_i; \hat{t}_i, m_i) \quad (8)$$

其中, θ_i 为图引导的注意力机制中的参数。门机制主要由融合门和过滤门组成。为获得文本和图片的多模态特征表示, 利用门机制中的融合门对新获得的依赖于 IGA 的词特征 \hat{m}_i 和依赖于 WGA 的图片特征向量 \hat{t}_i 进行拼接, 获得多模态融合后的中间特征表示。尽管利用 WGA 和 IGA 能够生成富含多模态语义特征的中间表示, 但是依然存在噪声。例如, 当预测文本中实体所包含的副词或形容词标签时, 与之对应的图片特征并不能提供语义表示的增强, 反而会引入噪声。因此, 应用门机制中的过滤门, 采用 Sigmoid 函数对融合后的多模态中间表示特征进行噪声过滤, 获得高质量多模态中间特征表示 g_i 。尽管融合后的多模态中间特征能够在一定程度上完成对文本和图片语义的联合表达, 但是命名实体识别的核心语义依旧在于文本。因此, 通过将 BiLSTM 编码后序列特征与多模态中间表示特征相拼接, 获得最终多模态表示特征 u_i , 具体过程为

$$g_i = \text{Gate}(\theta_g; \hat{t}_i, \hat{m}_i) \quad (9)$$

$$u_i = m_i \oplus g_i \quad (10)$$

其中, $g_i, u_i \in \mathbb{R}^d$; θ_g 为门机制中的全部参数。将编码的多模态特征 u_i 表示输入 CRF 层中进行标签解码, 并利用最大似然估计对 CRF 层进行训练, 获得解码标签。

3.4 深度迁移学习

为了缓解当前可利用标注语料匮乏的现状, 本文提出利用深度迁移学习方法探索解决这一问题, 并设计了基于深度迁移学习的 MNERM 模型。利用预训练完成的 BAC 模型和 ACAC 模型, 将相应的权重分别迁移至 BAM 模块和 ACAM 模块中, 使得 MNERM 模型具备对目标域 (地方志多模态数据集) 抽取文本和多模态特征的能力。

给定用于进行实体识别的地方志文本图片对 (C, P) , C 经过加载权重后的 BAM 模块得到的编码输出为 $C' = \{c_i | c_i \in \mathbb{R}^d, i = 1, 2, \dots, n\}$, (C, P) 经过加载权重后的 ACAM 模块得到的多模态特征表示 $K = \{k_i | k_i \in \mathbb{R}^d, i = 1, 2, \dots, n\}$ 。尽管利用迁移学习后的多模态特征能够在一定程度上增强文本语义信息, 但

是其仍包含一定的噪声,并且模型学习的语料并不是原始中文语料,而是英译汉平行语料,经过翻译后会部分丢失原意,引入噪声。因此,本文提出应用过滤门对提取的多模态特征进行噪声过滤,得到过滤后的多模态特征 $V = \{v_i | v_i \in \mathbb{R}^d, i = 1, 2, \dots, n\}$, 之后将文本语义特征表示 C' 与多模态特征表示 V 进行融合输入至一层全连接层中进行非线性激活,获得最终的多模态特征表示 $Z = \{z_i | z_i \in \mathbb{R}^{2d}, i = 1, 2, \dots, n\}$, 具体过程为

$$v_i = \text{Sigmoid}(W_k k_i + b_k) \quad (11)$$

$$z_i = \tanh(W_z(c_i \oplus v_i) + b_z) \quad (12)$$

其中, W_k 和 W_z 为权重矩阵; b_k 和 b_z 为偏置项; \tanh 为非线性激活函数。本文将多模态特征 Z 输入 CRF 层中, 微调后获得最终的预测标签。

4 实证研究

4.1 实验数据集

1) 人民日报语料库

本文使用的是1998年1月的人民日报语料库, 该语料库由北京大学计算语言研究所和富士通公司联合制作并发布, 被广泛应用在命名实体识别研究中。语料库中包含人名、地名及机构名实体, 本文以行对话料进行切分, 共获得19484条句子, 将语料库的80%作为训练集, 剩余的20%作为测试集。

2) 中文推特多模态数据集

本文使用的是Zhang等^[8]用于多模态命名实体任务的英文推特数据集。该数据集共包含8257个句子和图片对, 标注实体类别为人名、地名、机构名及其他实体, 利用BIO(begin, inside, outside)规则^[24]进行实体标注。该数据集经双人标注完成, 包含的实体数量为12784, 训练集句子数量为4000, 验证集数量为1000, 测试集数量为3257。为了制作平行语料, 本文首先利用科大讯飞翻译API(application programming interface)对数据集进行翻译, 并召集5位研究生对平行语料进行检查, 使其通顺并保持原意; 然后利用jieba包对语料进行分词, 并使用相同标注规则对照原英文语料进行实体标注; 最后得到中文推特多模态数据集。在英文推特中, 语料中常包含缩写词及非中文对应实体词, 同时考

虑到迁移应用的语料, 本文在中文平行数据集去除了其他实体类别。该平行数据集的训练集、验证集及测试集数量均与原数据集保持一致, 在实体对照的标注过程中, 当中文出现了英文中未标注的实体, 本文则加以补充, 最后得到的实体数量为10636。

3) 地方志多模态数据集

利用本课题组编写的爬虫对《南京简志》^①《南京人物志》^②《南京园林志》^③《南京城墙志》^④、百度中的南京地方志等资源进行爬取, 获取志书中的图片及相应文本描述, 文本均为现代文。搜集到的文本及图片对数量为2885, 经过过滤及去重, 共获得1659个文本图片对。之后对数据进行实体标注, 标注由组内的两位研究生完成, 标注规则为BIO^[24], 标注实体类别分别为人名、地名及机构名, 实体总量为2908。标注后的地方志多模态数据集作为检验本文提出的MNERM模型的性能测试语料。本文同时标注了500个用于微调的文本图片对。

4.2 实验设置

文本句子最大长度设置为150, 将多余的部分截断; 反之, 则补零。词向量的维度为300。在BAC模型中, 隐藏层单元数为200, dropout值为0.5, 训练批次大小为64, 训练轮数为100。在ACAC模型中, BiLSTM隐藏层单元数为200, dropout值为0.25, 训练批次大小为10, 训练轮数为100。BAC模型与ACAC模型所用的优化器为adam, 学习率为0.001。在MNERM模型中, 微调过程所用的优化器为adam, 学习率为0.001, 微调训练轮数为10, 同时BAM模块和ACAM模块中的权重在微调过程均被冻结。本文所用的评估指标分别为精确率(precision, P)、召回率(recall, R)及F1值。

本文所用编程语言为Python 3.6, 使用的深度学习框架为tensorflow2.3.0, 本文的实验均在两块GPU型号为NVIDIA GeForce RTX 2080ti、内存为16G的服务器中完成。

4.3 基线模型

基于深度迁移学习的MNERM模型主要由BAM模块、ACAM模块、过滤门及CRF层构成, 组成模块的性能影响着整体模型的表现。因此, 本文按照

① 江苏古籍出版社, 1986年出版。

② 学林出版社, 2001年出版。

③ 方志出版社, 1997年出版。

④ 凤凰出版社, 2008年出版。

使用的数据集,分别是人民日报语料库和中文推特数据集,将组成模块对应的模型(BAC和ACAC)与不同的基线模型进行对比,以验证其性能。最后,本文将MNERM模型在地方志多模态数据集进行性能验证,并与基线模型作对比。

1) 人民日报语料库

本文选择了几种具有优异性能的文本实体识别模型,将其与BAC模型作对比,具体如下。

BiLSTM-Att^[25]:该模型使用的注意力机制同BAC模型相同,解码层使用Softmax函数作为标签解码层。

BiLSTM-CRF^[26]:该模型结合了BiLSTM模型和CRF模型,具有良好的实体识别效果,并被广泛应用在命名实体识别任务中。

BiLSTM^[27]:相较于BiLSTM-CRF模型,该模型利用Softmax函数作为序列解码层,具有一定的实体识别性能。

CRF^[28]:该模型为命名实体识别任务中的经典模型,能够较好地考虑到序列特征并避免标签偏置问题。

2) 中文推特多模态数据集

ACAM模块主要由WGA、IGA和门机制组成,为验证组成部分的优越性,本文对基于ACAM的ACAC模型进行了消融实验,分别去除了WGA、IGA和门机制,形成Without-WGA、Without-IGA和Without-Gate等模型。同时,为了验证多模态融合的性能,本文将其与仅基于文本的BiLSTM-CRF作对比,具体如下。

Without-WGA:该模型去除了词引导的注意力机制,仅保留了图引导的注意力机制。

Without-IGA:该模型去除了图引导的注意力机制,仅保留了词引导的注意力机制。

Without-Gate:该模型在自适应联合注意力网络中去除了门机制。

BiLSTM-CRF^[27]:该模型对文本序列进行命名实体识别,参数与ACAC保持一致。

3) 地方志多模态测试数据集

为了验证MNERM模型的性能,本文将仅在人民日报语料库和中文推特数据集中进行预训练的BAC和ACAC作为对比模型,微调方式均保持一致。同时,为了验证过滤门的性能,本文设计了去除过滤门的模型Without-FGate作为对比。本文还将哈尔滨工业大学提供的Language Technology Platform (LTP)^[29]中的命名实体工具作为对比模型。

4.4 实验结果及分析

1) 人民日报语料库

表1呈现的是BAC模型与其他模型的对比结果。从表1可以看出,本文提出的模型在各个指标中均表现最优。在同BiLSTM-CRF的比较中可以发现,当模型的解码层均保持相同时,引入自注意力机制能够使模型更为关注那些能够决定序列标签的信息,生成富含语义特征的序列特征,从而提升模型识别实体的性能,这也是BAC模型具有一定优势的原因。在同BiLSTM-Att的对比中,当模型的编码层保持一致时,利用Softmax层作为识别实体的解码层,尽管能够取得一定的性能,但依旧劣于BAC模型。相较于Softmax层,CRF能够对隐藏层的各个时间步进行有效建模,学习并观察序列中的标签特点,从而提升模型的解码性能。这样的优势同样体现在BiLSTM和BiLSTM-CRF的对比中。当忽略文本的上下文关系,仅用词向量对文本进行表示时,将其输入CRF层进行解码,从结果可以发现,CRF模型均劣于使用BiLSTM或结合自注意力机制的模型作为上下文建模的模型,这充分说明了文本上下文在命名实体识别任务中的重要作用,同时也表明利用BiLSTM等时间序列模型能够较好地学习文本上下文关系,并能生成富含上下文关系及语义信息的序列特征。

表1 模型在人民日报语料库中的测试结果

模型	P(%)	R(%)	F1(%)
BiLSTM-CRF ^[26]	95.639	94.551	95.090
BiLSTM ^[27]	94.982	93.229	94.097
CRF ^[28]	88.429	88.248	88.336
BiLSTM-Att ^[25]	96.340	94.425	95.371
BAC	96.954	94.941	95.935

注:粗体表示在该指标下取得的最优结果。

通过比较分析发现,本文引入的BAC模型具有较好的实体识别性能,而模型包含的BiLSTM和自注意力网络在其中发挥了充分抽取语义特征的重要作用,这也是本文将BiLSTM和自注意力网络(BAM模块)作为MNERM模型组成部分的原因。

2) 中文推特数据集

自适应联合注意力机制由图引导的注意力机制、词引导的注意力机制及门机制组成。每个组成部分均能对ACAC模型性能产生影响,为了探究不同组成成分的作用及整体组合的性能,本文对此进行了探究。

表 2 呈现的是各对比模型在中文推特多模态数据集集中的结果,可以看出,ACAC 模型在 F1 这一指标上表现最优。当去除图引导的注意力机制后,Without-IGA 模型在精确率 (P) 这一指标上优于 ACAC 模型,但是在召回率 (R) 和 F1 指标上均劣于 ACAC。尽管 ACAC 模型在预测序列正标签样本中并没有表现出最优性能,但是在序列中的各实体类别真实标签样本识别中效果最佳,并在召回率这一指标上超出 With-IGA 模型近 7%。当去除词引导的注意力机制后,仅利用图引导的注意力机制并不能较好地学习到文本和图片之间的模态交互和关联关系,这也是 Without-WGA 劣于 ACAC 的原因。在同 Without-FGate 模型的对比中,可以发现门机制在模型中的重要作用,引入门机制能够较好地聚合多模态融合特征,同时有效过滤来自模态融合中的噪声。当不考虑文本对应的图片时,通过对比 BiLSTM-CRF,可以发现图片信息在增强文本语义特征中的作用,这也是 ACAC 模型表现良好的原因。因此,本文将去除了 CRF 层的 ACAC 模型作为 MNERM 模型中的 ACAM 模块,用于提取地方志数据中的多模态特征。

表 2 模型在中文推特数据集的测试结果

模型	$P(\%)$	$R(\%)$	$F1(\%)$
Without-IGA	68.876	59.636	63.924
Without-WGA	62.757	59.198	60.925
Without-FGate	64.501	62.241	63.351
BiLSTM-CRF ^[26]	63.834	55.579	59.421
ACAC	63.097	66.136	64.581

注:粗体表示在该指标下取得的最优结果。

3) 地方志多模态数据集

表 3 呈现的是经过微调后的不同对比模型对地方志多模态数据集进行实体识别的结果,各模型所用的微调数据均一致。利用通用模型 LTP 对地方志语料进行实体识别并没有取得较好的效果。与 BAC 模型比较可以发现,当 MNERM 模型联合多模态语料库知识后,模型性能有了较大提升。这表明在多模态语料库中预训练实体识别模型后,利用基于神经网络的深度迁移学习方法,将权重灌入 MNERM 模型对应模块中,能够使得 MNERM 具备捕捉不同模态间的语义相关关系及动态交互的能力,从而获得更优的性能。在与 ACAC 的比较中可以发现,尽管利用在中文推特多模态语料库中的预训练模型 ACAC 能够取得一定优势,但是劣于含有人民日报语料库知识的 BAC 模型以及 MNERM 模型。一方面

是因为在制作平行语料的过程中,会伴随着部分英文原意信息的丢失;另一方面是因为源域英文推特数据集大多由推特平台上用户的发帖组成,内容大多关于用户生活的分享,而目标域则是地方志多模态内容,目标域与源域之间存在着部分不相关的知识。当本文引入过滤门后可以看出,采用过滤门的 MNERM 模型在精确率和 F1 指标上均优于 Without-FGate 模型。尽管应用过滤门机制使得召回率轻微下降,但是 F1 值提升了 1.042%。这表明,应用过滤门能够对深度迁移学习得到的多模态融合特征噪声进行有效过滤,同时能够弥补因源域和目标域之间存在不匹配知识所造成的性能损失。

表 3 地方志多模态数据集深度迁移学习结果

模型	$P(\%)$	$R(\%)$	$F1(\%)$
LTP ^[29]	32.993	35.523	34.211
ACAC	55.921	57.097	56.316
BAC	74.796	70.086	72.288
Without-FGate	75.834	79.961	77.820
MNERM	78.710	79.077	78.862

注:粗体表示在该指标下取得的最优结果。

4) 深度迁移学习有效性分析

为了探究深度迁移学习在地方志多模态命名实体任务中的有效性以及模型对目标领域的适配性,本文通过调节预训练模型中训练集大小进行验证^[6]。图 3 展示的是当人民日报语料库训练集大小成比例增加时,BAC 模型在人民语料库中的测试性能及在地方志多模态数据集集中的文本进行深度迁移学习的结果。从图 3 可以看出,随着预训练模型中训练集数量的增加,经过微调后的权重迁移模型对地方志文本进行实体识别的性能呈上升趋势。该趋势同样呈现在 ACAC 模型对地方志多模态数据的实体识别中。

从图 4 可以看出,当人民日报语料库及中文推特多模态数据集集中的训练集同步成比例上升时,应用深度迁移学习的 MNERM 模型在对地方志多模态数据集集中的实体进行预测时,性能总体呈上升趋势。综合图 3、图 4 中的结果可以发现,预训练模型中训练集的大小影响着后续应用深度迁移学习的效果,这表明本文提出的深度迁移方法具有一定的有效性,并且显示出本文提出的 MNERM 模型对目标领域具有较强的适配性。

4.5 误差分析

表 4 呈现的是利用不同模型对地方志多模态数

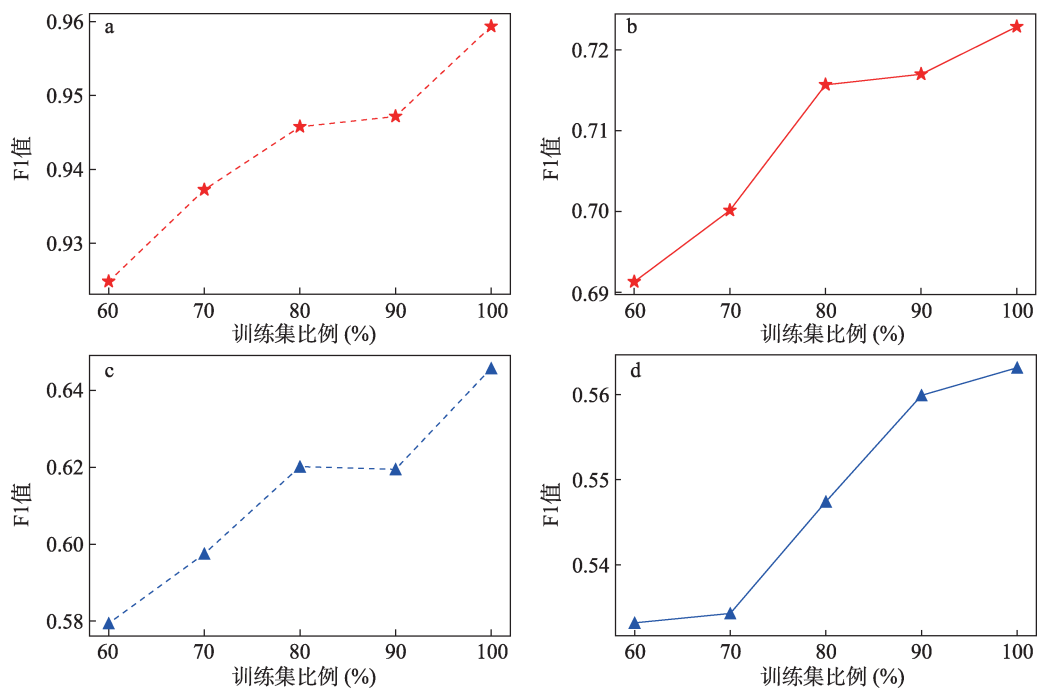


图3 训练集比例对BAC模型和ACAC模型性能及应用深度迁移学习的影响

a. BAC模型在人民日报语料库中的表现;b. 应用深度迁移学习后BAC模型在多模态地方志数据中的表现;
c. ACAC模型在中文多模态数据集集中的表现;d. 应用深度迁移学习后ACAC模型在多模态地方志数据中的表现。

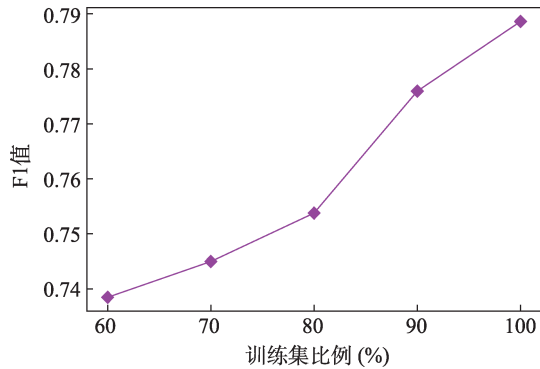




图4 预训练模型中的训练集比例对MNERM模型性能的影响

图中所示为MNERM模型在多模态地方志数据中的表现。

据集中的部分数据进行预测的结果。在例1中，MNERM模型和ACAC模型均对地名实体做出了准确的预测，而BAC模型则做出了错误判断。例1图片中的大楼为文本的地名实体提供了语义增强作用，通过多模态融合则可以产生更富含语义的表示，从而提升实体识别的性能。在多模态命名实体中，文本的语义信息依旧是实体识别的核心信息。在例2中，尽管利用ACAC模型未能对人名实体进行有效识别，但仅依靠文本语义信息，BAC模型做出了准确判断，而作为ACAC模型和BAC模型两者的结合，依靠捕捉文本语义信息的BAM模块，MNERM模型同样预测成功。在例3中，MNERM

表4 不同模型对地方志多模态数据进行实体识别的结果

示例			
<div><div></div><div></div><div></div></div>			
例1:[金陵饭店 LOC] ¹ 鸟瞰图 例2:军阀人物志系列——新桂系[李宗仁 PER] ¹ 例3:[倪国鼎 PER] ¹ ——[南京方志办 ORG] ² 供图			
MNERM模型	1=>LOC	1=>PER	1=>PER,2=>ORG
BAC模型	1=>未识别正确	1=>PER	1=>PER,2=>ORG
ACAC模型	1=>LOC	1=>未识别	1=>PER,2=>未识别

注:LOC表示地名实体,PER表示人名实体,ORG表示组织实体。

模型和BAC模型均对人名和组织实体做出了准确判断,而ACAC模型仅识别出了人名实体,未能识别出组织实体。例3图片中的人像为人名实体的识别提供了语义增强作用,但是在组织实体识别中,与文本相对应的图片未提供相应的补充特征,ACAC模型未能对组织实体进行识别。尽管MNERM模型在利用深度迁移学习的多模态命名实体识别任务中能够取得一定效果,但其未能够有效利用文本中的字级特征,而联合字级的特征则可以增强文本的表示能力,能够进一步改善多模态特征融合后的语义表示特征,从而提升迁移学习后实体识别的性能。

5 总结与展望

当前,面向地方志等文化资源的命名实体识别研究主要基于文本,忽略了文本对应的图片信息,同时还面临着在领域内训练实体识别模型缺乏已标注数据集的困境。为了解决该问题,本文从多模态视角出发,结合地方志对应的图片信息,并提出基于深度迁移学习的MNERM模型。该模型由四个部分组成,分别是BAM模块、ACAM模块、过滤门及CRF层。为了验证模型组成部分的有效性,本文将包含对应模块的模型(BAC和ACAC)与不同基线模型进行对比,实验结果表明,模型各组成部分均包含一定的优势。利用经过权重迁移后的BAM模块和ACAM模块,MNERM模型能够有效获取文本语义特征及多模态特征,应用过滤门对ACAM模块输出的多模态特征进行去噪,最后将BAM模块输出的文本语义特征及过滤后的多模态特征进行融合,输入至CRF层进行解码。实验结果表明,本文提出的模型在同基线模型的比对中具有一定优势。同时,为了验证深度迁移学习的有效性和对目标领域的适配性,本文将预训练模型中的训练集比例作为参数进行调节,发现当源域训练集越大,经过深度迁移学习后的模型表现越佳。

本文提出的模型和方法不仅适用于地方志多模态命名实体识别,也适用于数字人文领域中标注数据集匮乏的文化资源,如非遗等。在未来的研究中,本课题组将进一步提升模型的领域泛化能力,提升模型利用深度迁移学习进行多模态实体识别的性能以及中文多模态命名实体识别数据集的构建。

参 考 文 献

[1] 黄坤明. 推进社会主义文化强国建设[N]. 人民日报, 2020-

- 11-23(6).
- [2] 颜杰峰. 推进社会主义文化强国建设[J]. 红旗文稿, 2020(23): 34-36.
- [3] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.
- [4] 李娜. 基于条件随机场的方志古籍别名自动抽取模型构建[J]. 中文信息学报, 2018, 32(11): 41-48, 61.
- [5] 黄水清, 王东波, 何琳. 基于先秦语料库的古汉语地名自动识别模型构建研究[J]. 图书情报工作, 2015, 59(12): 135-140.
- [6] Lee J Y, Dernoncourt F, Szolovits P. Transfer learning for named-entity recognition with neural networks[C]// Proceedings of the Eleventh International Conference on Language Resources and Evaluation. European Language Resources Association, 2018: 4470-4473.
- [7] 王银瑞, 彭敦陆, 陈章, 等. Trans-NER: 一种迁移学习支持下的中文命名实体识别模型[J]. 小型微型计算机系统, 2019, 40(8): 1622-1626.
- [8] 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 Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 5674-5681.
- [9] 徐晨飞, 叶海影, 包平. 基于深度学习的方志物产资料实体自动识别模型构建研究[J]. 数据分析与知识发现, 2020, 4(8): 86-97.
- [10] 崔竞峰, 郑德俊, 王东波, 等. 基于深度学习模型的菊花古典诗词命名实体识别[J]. 情报理论与实践, 2020, 43(11): 150-155.
- [11] 唐慧慧, 王昊, 张紫玄, 等. 基于汉字标注的中文历史事件名抽取研究[J]. 数据分析与知识发现, 2018, 2(7): 89-100.
- [12] 殷章志, 李欣子, 黄德根, 等. 融合字词模型的中文命名实体识别研究[J]. 中文信息学报, 2019, 33(11): 95-100, 106.
- [13] 石春丹, 秦岭. 基于BGRU-CRF的中文命名实体识别方法[J]. 计算机科学, 2019, 46(9): 237-242.
- [14] 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.
- [15] 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.
- [16] Tan C Q, Sun F C, Kong T, et al. A survey on deep transfer learning[C]// Proceedings of the 27th International Conference on Artificial Neural Networks. Cham: Springer, 2018: 270-279.
- [17] 武惠, 吕立, 于碧辉. 基于迁移学习和BiLSTM-CRF的中文命名实体识别[J]. 小型微型计算机系统, 2019, 40(6): 1142-1147.
- [18] Li S, Zhao Z, Hu R F, et al. Analogical reasoning on Chinese morphological and semantic relations[C]// Proceedings of the 56th

- Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 138-143.
- [19] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
- [20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. (2015-04-10). <https://arxiv.org/pdf/1409.1556.pdf>.
- [21] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]// Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 248-255.
- [22] 孙凌浩. 利用翻译模型的跨语言中文命名实体识别[J]. 计算机工程与应用, 2021, 57(10): 94-100.
- [23] Carvalho T, de Rezende E R S, Alves M T P, et al. Exposing computer generated images by eye's region classification via transfer learning of VGG19 CNN[C]// Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications. IEEE, 2017: 866-870.
- [24] Tjong Kim Sang E F, Veenstra J. Representing text chunks[C]// Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1999: 173-179.
- [25] Rei M, Crichton G, Pyysalo S. Attending to characters in neural sequence labeling models[C]// Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. The COLING 2016 Organizing Committee, 2016: 309-318.
- [26] Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[OL]. (2015-08-09). <https://arxiv.org/pdf/1508.01991.pdf>.
- [27] Limsopatham N, Collier N. Bidirectional LSTM for named entity recognition in Twitter messages[C]// Proceedings of the 2nd Workshop on Noisy User-generated Text. The COLING 2016 Organizing Committee, 2016: 145-152.
- [28] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 2001: 282-289.
- [29] Che W X, Li Z H, Liu T. LTP: a Chinese language technology platform[C]// Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Stroudsburg: Association for Computational Linguistics, 2010: 13-16.

(责任编辑 王克平)