DOI: 10.16452/j.cnki.sdkjzk.2025.03.010 文章编号:1672-3767(2025)03-0097-10

一种基于并行多尺度特征学习的招聘信息抽取模型

郭雯靓[1],吕 楠[2],纪淑娟[1],步朝晖[3],王永政[1],曹 宁[1]

(1. 山东科技大学 山东省智慧矿山信息技术重点实验室,山东 青岛 266590;
2. 山东省慢性病医院(山东省康复中心),山东 青岛 266071;3. 青岛就业街大数据科技有限公司,山东 青岛 266555)

摘 要:随着网络招聘的普及,基于招聘广告的自动实体抽取,已成为职位和人才推荐等下游智能应用系统开发的重要基础。现有招聘广告实体抽取模型存在抽取内容分类覆盖不全面和超长文本序列语义稀释问题。本研究将招聘文本中的实体划分为四类,提出一种基于并行多尺度特征学习的招聘信息抽取模型(MUBLC)。首先,模型利用长短时记忆网络(LSTM)从原始数据中提取初始特征。然后,使用自注意力机制学习全局特征,采用动态深度卷积网络与自注意力共享投影的方式并行学习局部特征,同时在自注意力机制的每一层并行连接前馈神经网络,并行学习文本的逐点特征。最后,模型将并行学习得到的三种尺度特征进行向量融合,并输入条件随机场(CRF)获得预测的标签序列。实验结果表明,与现有最优模型相比,本研究所提模型的 F_1 值提高了 2.53%,表明并行学习三种特征能够有效缓解长序列语义稀释问题,显著提升招聘信息抽取性能。

关键词:网络招聘广告;招聘信息抽取;并行多尺度特征学习;动态深度卷积网络;命名实体识别

中图分类号:TP391.1 **文献标志码**:A

Recruitment information extraction model based on parallel multi-scale features learning

GUO Wenjing[1], LÜ Nan[2], JI Shujuan[1], BU Chaohui[3], WANG Yongzheng[1], CAO Ning[1]

(1. Shandong Key Laboratory of Wisdom Mine Information Technology, Shandong University of Science and Technology, Qingdao 266590, China;
2. Shandong Provincial Chronic Disease Hospital (Shandong Rehabilitation Center), Qingdao 266071, China;
3. Qingdao Employment Street Big Data Technology Co., Ltd., Qingdao 266555, China)

Abstract: With the popularity of online recruitment, automatic entity extraction based on job advertisements has become the basis of the development of downstream intelligent application systems such as job and talent recommendation. The existing entity extraction methods for job advertisements have problems such as incomplete coverage of extracted content classification and semantic dilution of ultra-long text sequences. In this paper, we divided the entities in the recruitment texts into four fine-grained categories, and proposed a recruitment information extraction model named multi-scale feature learning with bert-bilstm-crf (MUBLC) based on parallel learning of multi-scale features. Firstly, the model uses the long short-term memory (LSTM) network to extract initial features from the raw data. Secondly, the self-attention mechanism is used to learn global features. A dynamic deep convolutional network is used to learn local features in parallel with the self-attention sharing projection, and a feed-forward neural network is connected in parallel to each layer of the self-attention to learn the point-wise features of the text in parallel. Finally, the model fuses the three-scale features obtained through parallel learning into vectors and inputs them into the conditional random field (CRF) to obtain the predicted label sequence. Compared with the

收稿日期:2023-08-20

基金项目:国家自然科学基金项目(71772107);山东省自然科学基金项目(ZR2020MF044)

作者简介:郭雯靓(2001—),女,山东菏泽人,硕士研究生,主要从事知识图谱、推荐系统的研究.
纪淑娟(1977—),女,山东青岛人,教授,博士生导师,研究方向为人工智能、智能信息处理,本文通信作者.
E-mail:jsjsuzie@sina.com

state-of-the-art baseline model, the F_1 score of the proposed model is improved by 2.53%. The experimental results show that parallel learning of the three features is of great significance to solve the problem of long sequence semantic dilution, and it is feasible and effective to improve the performance of recruitment information extraction.

Key words: online recruitment advertisement; recruitment information extraction; parallel multi-scale features learning; dynamic deep convolutional network; named entity recognition

随着互联网的普及,网络招聘已经成为招聘的主流方式和求职者的重要途径。网络招聘信息呈现半结构化特征,既包含岗位名称、学历、薪资等结构化数据,又包含岗位职责描述等非结构化文本。招聘信息抽取是指从非结构化文本中抽取结构化招聘实体信息的过程,如从岗位职责描述文本中抽取专业技能、能力素质等信息。招聘信息抽取不仅可以为求职者了解岗位特征与技能需求提供参考,还可以为企业招聘和岗位精准推荐提供数据支撑[1]。

招聘信息抽取可视为一个命名实体识别任务。近年来,基于深度学习的命名实体识别在通用领域取得了显著进展,但针对招聘领域的命名实体识别研究相对较少。现有招聘领域的命名实体识别研究[2-12]面临两大挑战:一是缺乏对招聘领域实体的细粒度分类标注,导致抽取内容不完整,难以满足下游职位和人才推荐等智能应用对单位需求的精确挖掘;二是网络招聘文本普遍较长(通常超过 400 字),而传统命名实体识别方法在处理此类长文本时易出现语义稀释问题,导致识别效果不佳,难以直接应用于招聘领域。

为了提升招聘实体抽取性能并精确划分实体类型,本研究提出一种适用于长文本序列的招聘信息抽取模型。主要贡献有:

1) 构建并标注了一个招聘信息抽取数据集,并将招聘文本中的实体划分为四类,以解决现有方法抽取内容不完整和不规范等问题。这四类实体涵盖了招聘广告文本中的岗位职责、能力、经验和技能等关键信息,为细粒度挖掘招聘单位的人才需求提供基础。

2) 提出一种基于并行多尺度特征学习的招聘信息抽取模型(multi-scale feature learning with bert-bil-stm-crf,MUBLC),解决了传统模型仅提取全局特征导致的语义稀释问题。与现有缓解双向长短时记忆网络(bidirectional long short-termmemory,Bi-LSTM)长文本语义稀释问题的特征学习方法相比,MUBLC 模型在并行学习全局和局部特征的同时,还学习了逐点特征。

3) MUBLC 模型在招聘信息抽取任务中效果优于其他模型,验证了三种尺度特征并行学习在提升长文本实体识别性能方面的有效性。

1 相关工作

1.1 招聘信息抽取

根据抽取方法的不同,招聘信息抽取可分为手工抽取和自动抽取两大类。表 1 总结了现有招聘信息抽取相关研究。可以看出,现有研究主要集中于专业技能词的抽取[2-8,12],少量研究关注了招聘单位对求职者

表 1 招聘信息抽取模型的相关工作

Table 1 Comparison of works related to recruitment information extraction

抽取类型		模型	抽取对象	缺点
基于手工的模型		文献[2-3]	专业技能	耗费人力,难以满足招聘大数据环境
自动模型	基于外部资源的模型	文献[4-6]	专业技能	外部资源更新慢、覆盖面窄
	基于规则的模型	文献[7-8]	专业技能	方法简单、抽取效果不佳
	基于统计的模型	文献[9] 文献[10]	专业技能、经验要求 专业技能、能力素质	特征选取困难,语料库依赖大
	基于深度学习的模型	文献[11] 文献[12]	专业技能、能力素质等 专业技能	无法高效捕捉招聘领域文本特征

的其他要求,如经验与能力[9-11]。这些研究主要关注求职者应具备的技能,而忽略了他们需要“做什么”,即岗位职责。这些未被抽取的实体信息对于理解岗位需求、实现精准发现和推荐至关重要。此外,目前的招聘信息抽取方法大多基于手工和外部资源的方法,不仅耗费人力,而且难以适应大数据时代的需求。虽然基于深度学习的招聘信息抽取方法取得了较好的效果,但网络招聘广告文本普遍较长(400 字以上),在使用 Bi-LSTM 网络进行特征提取时,由于生成特征向量固定,易导致长序列语义稀释,影响抽取性能。

1.2 通用领域命名实体识别模型

作为文本信息自动抽取的核心技术,命名实体识别一直是文本信息处理领域的研究热点和难点。Li 等[13]提出一种改进的 Transformer 模型 NFLAT(non-flat-lattice transformer),该模型采用先“字-词”后“字-字”级别的自注意力机制,融合词边界和语义信息以提升字符表征能力。Wu 等[14]将汉字特征和部首嵌入相结合,提出一种使用双流 Transformer 结构的 MECT(multi-metadata embedding based cross-transformer)模型。除 Bi-LSTM 外,许多学者探索了命名实体识别的其他特征学习方法。例如,Liu 等[15]采用了 Bi-LSTM 的变体双向门控循环单元(bidirectional gated recurrent units,Bi-GRU)作为特征学习模块,其优势在于能够提高模型的运行速度;Zhu 等[16]结合 Bi-GRU 和卷积神经网络(convolutional neural networks,CNN)的优势,使用 Bi-GRU 学习全局特征,CNN 学习局部特征;蒋翔等[17]使用 CNN 的变体空洞卷积神经网络(iterated dilated convolutional neural network,IDCNN)提取文本特征,增强了模型识别上下文局部特征的能力。

这些全局或局部特征学习方法在一定程度上提升了实体识别效果,但缺乏针对超长序列文本实体识别的专门研究,难以有效解决变长和超长文本长序列的权重分散和语义稀释问题,因此在招聘信息抽取方面表现不佳。

2 基于并行多尺度特征学习的招聘信息抽取模型

本研究提出的基于并行多尺度特征学习的招聘信息抽取模型 MUBLC 总体框架如图 1 所示。该模型由五部分组成:嵌入层、初始特征提取层、并行多尺度特征学习层、多尺度特征融合层和解码层。首先,利用 BERT(bidirectional encoder representations from transformers)预训练模型将输入的招聘文本 T 编码为向量表示 $\boldsymbol{E}$;然后,将 $\boldsymbol{E}$ 输入Bi-LSTM中提取初始文本特征向量 $\boldsymbol{H}$,再将 $\boldsymbol{H}$ 输入并行多尺度特征学习层,

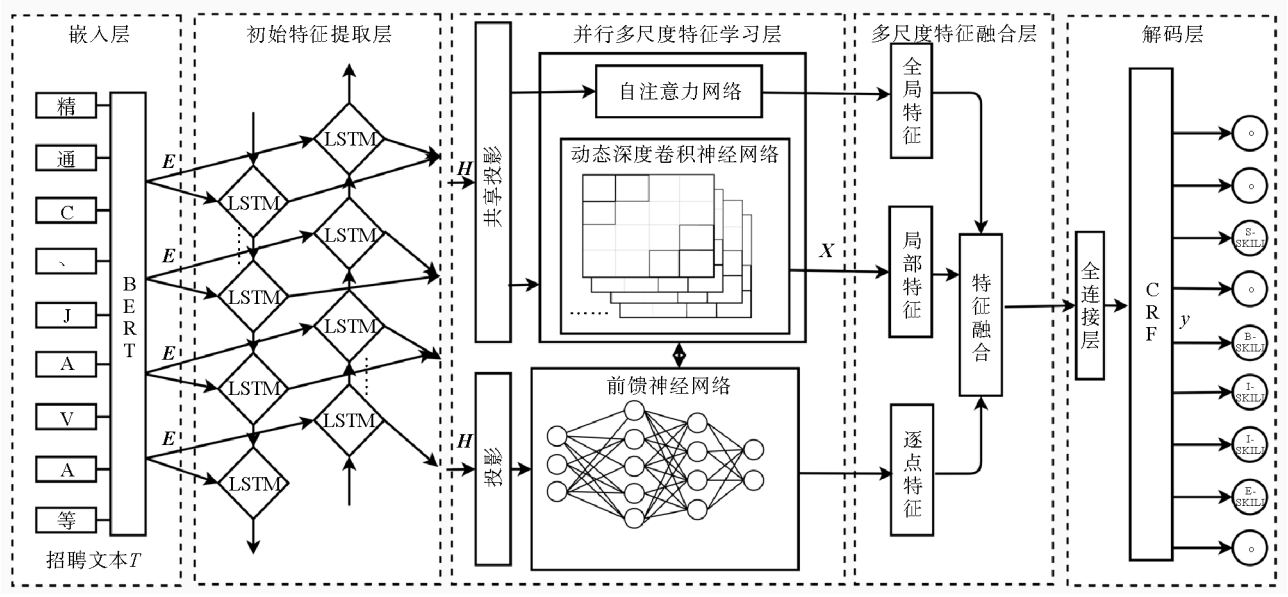

图 1 MUBLC 模型的总体框架

Fig. 1 Framework of MUBLC

通过自注意力网络、动态深度卷积神经网络和前馈神经网络并行学习文本的全局特征、局部特征和逐点特征;之后,将学习到的三种尺度特征进行向量融合,得到多尺度特征向量 $\boldsymbol{X}$;最后,将多尺度特征 $\boldsymbol{X}$ 输入全连接层,再输入到条件随机场(conditional random field,CRF)层中计算最优标签序列,得到最终的实体预测结果。

2.1 文本嵌入

嵌入层中,利用 BERT 预训练模型将输入的招聘文本转换为计算机可识别的向量。具体来说,初始文本 T 输入后,BERT 在每条句子首部插入[CLS]符号,在句子末尾插入[SEP]符号作为句子间的分割符,并将文本转换为字向量、文本向量、位置向量作为输入,最终将三类向量的加和作为输出。其中,字向量用于编码文本中的每个字符,文本向量用于区分句子并编码句子的语义信息,位置向量用于保证文本序列的顺序性。

2.2 初始特征提取

招聘信息实体特征的识别受到前后文语义的影响,因此,模型采用 Bi-LSTM 网络同时学习文本的历史和未来信息,以提取初始文本特征。随着序列长度的增加,传统循环神经网络(recurrent neural network,RNN)出现长距离依赖能力差、梯度消失或爆炸等问题。LSTM 作为一种特殊的 RNN,在隐藏层 h 中加入了三个门控结构,缓解了长距离依赖问题。三个门结构分别为:

$$i_t=\sigma(W_i\cdot[x_t,h_{t-1},C_{t-1}]+b_i), \tag{1}$$

$$f_t=\sigma(W_f\cdot[x_t,h_{t-1},C_{t-1}]+b_f), \tag{2}$$

$$o_t=\sigma(W_o\cdot[x_t,h_{t-1},C_{t-1}]+b_o)。 \tag{3}$$

式中:i_t、f_t、o_t 分别为输入门、遗忘门和输出门;σ 为 Sigmoid 激活函数;$\boldsymbol{W}=(w_i,w_f,w_o)$为权重向量;$\boldsymbol{b}=(b_i,b_f,b_o)$为偏置向量;$C_{t-1}$ 为上一时刻 LSTM 的细胞状态,此刻的细胞状态 C_t 按照式(4)进行更新,

$$C_t=f_t\cdot C_{t-1}+i_t\cdot\tanh(W_c\cdot[x_t,h_{t-1}]+b_c), \tag{4}$$

t 时间隐藏层 h_t 的状态依据式(5)进行更新,

$$h_t=o_t\cdot\tanh C_t。 \tag{5}$$

Bi-LSTM 由正向和逆向 LSTM 单元组成。将 BERT 编码后的向量 $\boldsymbol{E}$ 输入 Bi-LSTM 后,分别以正序和逆序进行特征提取,并将二者的输出向量拼接形成的特征向量 $\boldsymbol{H}$,从而使特征向量同时具有上下文信息。

2.3 并行多尺度特征学习

并行多尺度特征学习层由三个模块组成,分别为基于自注意力网络的全局特征学习模块、基于动态深度卷积网络的局部特征学习模块和基于位置前馈神经网络的逐点特征学习模块。模型将上一层 Bi-LSTM 生成的特征向量输入至该层并行提取文本的全局、局部与逐点特征。

本层通过自注意力网络学习全局序列的表征。注意力网络的本质是通过计算词与词之间的关联程度,获取词语结构信息,对文本中字符进行权重分配。采用自注意力网络可以提高重点词的权重,同时减少无用词对实体识别效果的影响。对于给定的输入 $\boldsymbol{H}$,首先,将其投射到三个表示键 K、查询 Q 和值 V 中,然后,利用自注意力网络学习全局特征 Attention($\boldsymbol{H}$),如式(6)所示:

$$\text{Attention}(\boldsymbol{H})=\gamma(QW^Q,KW^K,VW^V)W^O。 \tag{6}$$

式中:W^Q、W^K、W^V 为投影参数,W^O 为线性变换参数,γ 表示自注意力网络生成 K、Q、V 的操作,

$$\gamma(Q_1,K_1,V_1)=\text{softmax}(\frac{Q_1K_1}{\sqrt{d_k}})V_1。 \tag{7}$$

式中:V_1 进行了投影运算,即 $V_1=VW^V$。

自注意力网络存在权重过于集中或分散的问题,难以有效支持长序列表征学习。具体而言,自注意力网络的性能会随着文本序列长度的增加而下降,这是由于自注意力网络需要为所有位置分配全局权重,而长序列文本中权重可能过度集中于局部或发散至无关位置,导致仅能捕获文本的片段化特征。对于较短的序列,自注意力网络具有较好的性能,但对于较长的序列,会导致信息表示不足,进而阻碍模型对完整理解源信息的理解。因此,模型需要对长序列文本进行局部特征学习。

已有研究[16-17]采用 CNN 及其变体实现局部特征学习。然而,CNN 难以和自注意力网络在同一隐空间内并行提取特征。不同于 CNN 不可分离的卷积算子,本层选用动态深度卷积网络(Depthconv),通过共享自注意力机制的投影矩阵,实现与自注意力网络的结构兼容性。通过此设计,模型可并行学习局部上下文与全局依赖关系,显著提升长序列语义建模能力。

在动态深度卷积网络中,每个卷积子模块包含多个核大小不同的单元,可以用来捕捉不同范围的特征。核大小为 k 的卷积单元$\mathrm{conv}_k(\boldsymbol{H})$的输出为:

$$\mathrm{conv}_k(\boldsymbol{H})=\mathrm{Depthconv}_k(V_2)W^{\mathrm{out}}。\tag{8}$$

式中:W^{out} 为投影参数,Depthconv 为动态深度卷积单元的运算。在深度卷积网络模块中,其输入 V_2 进行了与自注意力机制相同的投影运算,即 $V_2=VW^V$。

对于输入序列 $\boldsymbol{H}$,动态深度卷积网络的输出 $\boldsymbol{O}$ 的计算式为:

$$O_{i,c}=\mathrm{Depthconv}_k(\boldsymbol{H})=\sum_{j=1}^{k}\left(\mathrm{softmax}\left(\sum_{c=1}^{d}W_{j,c}^{Q}H_{i,c}\right)\cdot H_{i+j-\frac{k+1}{2},c}\right)。\tag{9}$$

式中:i 表示输入特征向量的第 i 个元素,c 为动态深度卷积网络的输出通道,k 为卷积核的大小,d 为隐藏层的大小。

为了在同一隐空间并行学习全局与局部上下文序列表示,自注意力机制的投影 $V_1=VW^V$ 和动态深度卷积机制中的投影 $V_2=VW^V$ 是共享的,通过共享投影可以将输入特征投影到相同的隐空间中。

为了动态选择卷积核,本研究使用门控机制来自动选择不同卷积单元的权重,学习到的局部特征 Conv(**H**)如式(10)所示:

$$\mathrm{Conv}(\boldsymbol{H})=\sum_{i=1}^{n}\frac{\exp(\alpha_i)}{\sum_{j=1}^{n}\exp(\alpha_j)}\mathrm{conv}_{k_i}(\boldsymbol{H})。\tag{10}$$

式中:$\mathrm{conv}_{k_i}(\boldsymbol{H})$表示核大小为 k 的卷积单元,α 为卷积单元的权重,n 为卷积单元的数量。

招聘领域的实体识别存在较多单个字符实体,且存在许多实体与实体之间间隔字符为 1 的情况,如“C、C++语言”和“excel 等办公软件”。因此,提取文本的字符级特征是非常有必要的。

为了同自注意力网络和深度卷积网络并行学习逐点特征,模型将自注意力网络与前馈神经网络并行连接起来。由于线性变换在不同位置上是相同的,因此可将前馈神经网络视为一个字符级特征提取器,学习到的逐点特征 Pointwise(**H**)为:

$$\mathrm{Pointwise}(\boldsymbol{H})=\max(0,\boldsymbol{H}w_1+\lambda_1)w_2+\lambda_2。\tag{11}$$

式中:w_1、λ_1、w_2、λ_2 为投影参数。

2.4 多尺度特征融合

多尺度特征融合层将上一层生成的三组特征以向量融合的方式输出表示为 $\boldsymbol{X}$,

$$\boldsymbol{X}=\mathrm{Attention}(\boldsymbol{H})+\mathrm{Conv}(\boldsymbol{H})+\mathrm{Pointwise}(\boldsymbol{H})。\tag{12}$$

通过多尺度特征融合层,模型在兼顾捕获全局与局部特征的同时,还可以捕获潜在的逐点特征。

2.5 解码

招聘信息抽取任务需对文本进行序列标注。模型通过文本嵌入、初始特征提取及并行多尺度特征学习与融合后,最终引入条件随机场(CRF)[18]对输出序列标签进行约束。无约束的标注可能输出错误的标签,而 CRF 能从训练数据中获得约束性规则,为最后预测的标签添加约束,保证预测标签合法性,以得到全局最优标签序列。

假设输入文本 T 对应的输出标签序列为 y,将经过并行多尺度特征融合层得到的特征融合向量 $\boldsymbol{X}$,输入 CRF 中计算每个标签的转移概率,对应标签序列 y 的概率分布为:

$$P(y/\boldsymbol{X})=\frac{\exp(s(\boldsymbol{X},y))}{\sum_{y\in Y_x}\exp(s(\boldsymbol{X},y))}。\tag{13}$$

式中:$s(\boldsymbol{X},y)=\sum_{i=1}^{n}(W_{y_i,y_{i+1}}+P_{i+1,y_{i+1}})$为所有潜在的标签序列,模型最终输出结果为标签概率分布最大的标签序列,$s(\boldsymbol{X},y)$表示每个标签序列对应的得分,计算式为:

$$s(\boldsymbol{X},y)=\sum_{i=1}^{n}(W_{y_i,y_{i+1}}+P_{i+1,y_{i+1}})。\tag{14}$$

式中:$\boldsymbol{W}$ 为转移矩阵,$W_{y_i,y_{i+1}}$ 为标签从 y_i 转移到 y_{i+1} 的得分,$P_{i+1,y_{i+1}}$ 为第 $i+1$ 个字对应标签的得分。

3 实验设置与结果分析

3.1 数据集构建

本研究从前程无忧网站爬取互联网相关岗位招聘数据作为初始数据,提取其中的岗位描述文本进行标注。由于时间与标注成本限制,本研究采用 BIEOS 标注方式,标注了 2 800 条岗位描述文本。为了能够涵盖大多数的岗位描述信息,本研究将招聘信息分为 4 类实体,其类别及示例如表 2 所示。

表 2 实体标注类别及示例

Table 2 Type and sample of entity annotation

实体标签	实体含义	实体解释	实体示例
DUTY	工作职责	该岗位所需完成的工作	界面开发、代码优化、软件性能测试
QUA	能力素质	胜任该岗位所需的个人品质或能力	吃苦耐劳、责任心、团队合作精神
EXP	经验要求	胜任该岗位所需实践经验或相关经历	移动端开发经验、管理经验、留学经历
SKILL	专业技能	胜任该岗位所需专业或具备的技能	人工智能、Python、办公软件、电子商务

数据集的组成情况如表 3 所示,并按 8∶1∶1 的比例划分为训练集、验证集和测试集。每条招聘岗位文本平均含有 447 个字符和 16 个岗位描述实体。其中,描述工作职责的实体 6～7 个,描述能力素质的实体 5～6 个,描述经验要求的实体 1 个,描述专业技能的实体 2～3 个。

3.2 实验设置

表 3 实体个数统计情况

Table 3 Statistics of the number of entities

数据集	DUTY	QUA	EXP	SKILL
训练集	15 053	12 358	2 521	5 678
验证集	1 933	1 546	374	866
测试集	1 999	1 276	298	597
实体总数	18 985	15 180	3 193	7 141
每条岗位文本的平均实体数	6.8	5.4	1.1	2.6

3.2.1 参数设置

MUBLC 模型在训练过程中采取固定 BERT 参数方式,学习率为 5e-5,batch sizes 为 16,epoch 设为 30,dropout 为 0.2,最大序列长度设置为 512 个字符,当文本长度大于 512 字符时,采取从左向右截取 512 个字符作为模型输入,若长度不足 512 字符,则使用 padding 方法将输入文本补足。

3.2.2 对比实验设置

为验证 MUBLC 模型的性能,与 Bi-LSTM[19]、BiLSTM-CRF[20]、BERT-CRF[21]、Lattice-LSTM[22]、LR-CNN[23]、NFLAT[13]、MECT[14]、BERT-BiLSTM-CRF[24]、BERT-BiGRU-CRF[15]、CAN-NER[16] 和 BiLSTM-IDCNN-CRF[17] 模型进行对比实验。

Bi-LSTM[19] 将文本输入 Word2Vec 中进行编码,将向量输入 Bi-LSTM 中进行训练并输出预测标签序列。BiLSTM-CRF[20] 是命名实体识别经典模型,将 Word2Vec 的编码向量输入到 Bi-LSTM 模型中进行训练,最后通过 CRF 解码层优化并输出预测标签序列。BERT-CRF[21] 使用 BERT 预训练模型对输入文本进行向量编码,将向量输入至 CRF 解码层得到预测标签序列。Lattice-LSTM[22] 在 LSTM 基础上增加存储词汇的结构,通过门控循环单元利用字符序列的词汇信息缓解分词错误。LR-CNN[23] 基于卷积神经网络并行地对所有匹配句子的潜在单词进行建模。NFLAT[13] 先进行“字-词”级别的自注意力机制,从而获得融合了词边界和语义信息的字符表征,再进行“字-字”级别的自注意力机制,生成最终文本表征。MECT[14] 使用双

流 Transformer 将汉字特征和部首嵌入相结合。BERT-BiLSTM-CRF[24]使用 Bi-LSTM 学习文本全局上下文特征。BERT-BiGRU-CRF[15]采用具有全局注意力机制的 Bi-GRU 学习文本全局特征。CAN-NER[16]使用具有局部注意力机制的 CNN 和具有全局注意力机制的 Bi-GRU 分别学习文本的局部和全局特征。BiLSTM-IDCNN-CRF[17]使用 Bi-LSTM 提取初始全局特征,然后输入 IDCNN 中进一步学习局部特征。其中,Bi-LSTM、BiLSTM-CRF 和 BERT-CRF 是命名实体识别领域普遍采用的模型,Lattice-LSTM、LR-CNN、NFLAT、MECT 为命名实体识别领域的先进模型,因此将这七种模型作为基准模型。后四种模型采用四种不同特征学习方式,用来验证 MUBLC 模型并行融合不同尺度特征的有效性。

3.2.3 消融实验设计

为证明模型中各个模块的必要性,特别是并行提取逐点特征、局部特征和全局特征模块对招聘信息抽取性能提升的有效性,设计了六组消融实验,如表 4 所示。

表 4 消融实验设置

Table 4 Setup of the ablation experiment

模型	BERT 文本嵌入	初始特征提取	全局特征学习	局部特征学习	逐点特征学习
BERT-MU-CRF	√	—	√	√	√
BiLSTM-MU-CRF	—	√	√	√	√
BERT-BiLSTM-CRF[24]	√	√	—	—	—
ATTBLC	√	√	√	—	—
MUBLC-l	√	√	√	—	√
MUBLC-p	√	√	√	√	—
MUBLC	√	√	√	√	√

3.2.4 评价指标

本研究选择精确率(P)、召回率(R)与 F_1 值作为模型的评价指标,评价指标数值越高,表明模型的性能越好。

3.3 实验结果与分析

3.3.1 对比实验结果与分析

MUBLC 分别与基线模型和四种不同特征学习方法的模型的对比实验结果如表 5 和表 6 所示。MUBLC 模型对比性能最好的 NFLAT 模型,其 F_1 值提高了 2.53%。实验结果表明,MUBLC 模型在招聘信息抽取任务上优于其他基线模型。由表 6 可以看出,相比于其他特征提取方法,本研究提出的并行多尺度特征学习和融合方法取得最好的结果。实验结果表明,本研究提出的并行多尺度特征学习方法在招聘信息抽取方面优于其他特征学习方法,通过对文本的三种尺度特征进行并行学习,捕捉长短不一序列的文本特征,增加重点字词句的权重,能够有效提升招聘信息的实体识别性能。

3.3.2 不同类型实体抽取性能对比

为进一步验证各类岗位描述实体的抽取效果,本研究分析了模型在不同招聘信息实体类型上的识别性能,如表 7 所示。由表 7 可以看出,MUBLC 模型对 EXP 类实体的识别效果提升最大,相较于最先进的基线模型 CAN-NER,F_1 值提升了 3.13%;模型对 QUA 类实体提升最小,相比于 CAN-NER 模型,F_1 值仅提升 1.31%,这是因为 QUA 类实体长度变化较小,大多固定在 3～4 字。四类实体中,QUA 类实体识别效果最好,因为描述能力与素质的词语相对正式和固定。但 DUTY 和 SKILL 这两类实体识别精确率一般,是因为这两类实体的长度较长、边界很难识别,并且 DUTY 类实体往往由于岗位类型的多样性或招聘人员不同的写作风格而差别较大,而 SKILL 类实体往往包含各种实体嵌套和缩略词,导致实体难以预测。

表 5 基线模型对比实验

Table 5 Comparison experiments with baseline models %

模型	P	R	F_1
Bi-LSTM[19]	72.37	68.27	70.26
BiLSTM-CRF[20]	74.92	69.81	72.27
BERT-CRF[21]	80.36	79.73	80.00
Lattice-LSTM[22]	74.72	69.42	71.98
LR-CNN[23]	74.70	81.69	78.04
NFLAT[13]	78.03	84.84	81.20
MECT[14]	83.35	78.14	80.66
MUBLC	82.74	84.73	**83.73**

表 6 采用不同特征学习方法的模型对比实验

Table 6 Comparison experiments using different feature extraction methods %

模型	P	R	F_1
BERT-BiLSTM-CRF[24]	79.30	83.12	81.16
BERT-BiGRU-CRF[15]	74.70	76.88	75.78
CAN-NER[16]	79.50	81.48	80.48
BiLSTM-IDCNN-CRF[17]	78.89	82.74	80.76
MUBLC	**82.74**	**84.73**	**83.73**

表 7 不同实体类型抽取性能结果

Table 7 Performance experiments for different entity categories %

模型	DUTY			QUA			EXP			SKILL		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Lattice-LSTM[22]	63.55	56.41	59.77	88.70	85.39	87.01	77.73	77.56	77.64	72.62	67.62	70.03
BERT-BiGRU-CRF[15]	65.98	68.30	67.12	86.51	91.53	88.95	75.60	71.01	73.23	72.49	72.71	72.60
CAN-NER[16]	71.63	74.73	73.14	90.52	93.58	92.02	80.60	84.89	82.68	78.29	75.02	76.62
MUBLC	**75.44**	**78.12**	**76.76**	**92.40**	**94.26**	**93.33**	**85.62**	**86.00**	**85.81**	**82.23**	**83.38**	**82.80**

3.3.3 消融实验结果与分析

本研究设计了模型的六个变体进行消融实验,不同模块对模型性能的影响结果如表 8 所示。由表 8 可以看出,在不同模块对 MUBLC 模型的性能影响中,BiLSTM-MU-CRF 模型由于失去了 BERT 编码器的精确向量表示,极大影响模型捕获文本特征的能力,其 F_1 值相较 MUBLC 下降了 11.6%。对于多尺度特征学习模块,局部特征学习对模型影响较大,在全局特征学习的基础上增加局部特征学习时,F_1 值提高了 1.39%;逐点特征学习模块对模型影响较小,在全局特征学习的基础上增加逐点特征学习,F_1 值仅提高了 0.01%,但在全局特征和局部特征学习的基础上增加逐点特征学习模块时,F_1 值提高了 0.65%。实验结果表明,在招聘信息抽取任务中,仅学习逐点特征而不学习局部特征往往不能有效捕捉实体词之间的权重特征。

表 8 消融实验结果

Table 8 Results of ablation experiments %

模型	P	R	F_1
BERT-MU-CRF	83.19	80.20	81.66
BiLSTM-MU-CRF	74.81	69.64	72.13
BERT-BiLSTM-CRF	79.30	83.12	81.16
ATTBLC	78.87	84.72	81.69
MUBLC-l	79.31	84.24	81.70
MUBLC-p	81.50	84.73	83.08
MUBLC	82.74	84.73	83.73

3.3.4 对长序列表征的有效性分析

自注意力网络的性能会随着文本长度的增加而下降,为验证并行多尺度特征学习对长序列表征的有效性,本研究分别选取 500 条文本长度大于 100 字符与小于 100 字符的招聘样本进行实验,不同模型的 F_1 值对比结果如表 9 所示,其中 BLC 是基线模型 BERT-BiLSTM-CRF。当序列长度小于 100 字符时,ATTBLC 和 MUBLC 的 F_1 值相较基线模型分别提升了 1.54%和 1.93%;当序列长度大于 100 字

表 9 不同长度招聘文本下的不同模型的 F_1 值对比

Table 9 F_1 score of different models under different text lengths %

文本长度	BLC[24]	ATTBLC	MUBLC
<100	84.63	86.17	**86.56**
100~512	80.96	81.16	**83.53**

符时,MUBLC 相比于基线模型提高了 2.57%,而仅采用自注意力网络提升了 0.2%。实验结果表明并行多尺度特征学习在长序列表征中的有效性。

4 结论

本研究针对现有招聘信息抽取研究中存在的抽取信息不完整的问题,规范了招聘文本的抽取信息类型,构建了招聘信息抽取数据集,并将招聘信息分为四类实体。针对已有模型的局限性和网络招聘广告文本特征,提出了基于并行多尺度特征学习的招聘信息抽取模型 MUBLC。实验结果表明,MUBLC 模型在招聘领域的长文本特征抽取上优于现有基线模型,验证了该模型的有效性。

参考文献:

[1] 岳铁骐,傅友斐,徐健.基于招聘广告的岗位人才需求分析框架构建与实证研究[J].数据分析与知识发现,2022,6(2/3):151-166.
YUE Tieqi,FU Youfei,XU Jian. An analysis framework for job demands from job postings[J]. Data Analysis and Knowledge Discovery,2022,6(2/3):151-166.

[2] KIM J Y,LEE C K. An empirical analysis of requirements for data scientists using online job postings[J]. International Journal of Software Engineering and Its Applications,2016,10(4):161-172.

[3] DE MAURO A,GRECO M,GRIMALDI M,et al. Beyond data scientists:A review of big data skills and job families[C]//Proceedings of International Forum on Knowledge Asset Dynamics. 2016:1844-1857.

[4] ZHAO M,JAVED F,JACOB F,et al. SKILL:A system for skill identification and normalization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2015,29(2):4012-4017.

[5] 詹川.基于文本挖掘的专业人才技能需求分析:以电子商务专业为例[J].图书馆论坛,2017,37(5):116-123.
ZHAN Chuan. Analysis of professionals' skill demands based on text mining:Illustrated by the case of e-commerce[J]. Library Tribune,2017,37(5):116-123.

[6] 夏立新,楚林,王忠义,等.基于网络文本挖掘的就业知识需求关系构建[J].图书情报知识,2016,169(1):94-100.
XIA Lixin,CHU Lin,WANG Zhongyi,et al. The establishment of demand relationship of employment knowledge based on Web text mining[J]. Documentation,Information and Knowledge,2016,169(1):94-100.

[7] BASTIAN M,HAYES M,VAUGHAN W,et al. LinkedIn skills:Large-scale topic extraction and inference[C]//Proceedings of the 8th ACM Conference on Recommender systems. 2014:1-8.

[8] 王召义,薛晨杰,刘玉林.基于邻近词分析的电子商务技能需求分析[J].信息资源管理学报,2018,8(2):113-121.
WANG Zhaoyi,XUE Chenjie,LIU Yulin. The analysis of e-commerce skills based on the analysis of adjacent words[J]. Journal of Information Resources Management,2018,8(2):113-121.

[9] 刘睿伦,叶文豪,高瑞卿,等.基于大数据岗位需求的文本聚类研究[J].数据分析与知识发现,2017,1(12):32-40.
LIU Ruilun,YE Wenhao,GAO Ruiqing,et al. Research on text clustering based on requirements of big data jobs[J]. Data Analysis and Knowledge Discovery,2017,1(12):32-40.

[10] CALANCA F,SAYFULLINA L,MINKUS L,et al. Responsible team players wanted:An analysis of soft skill requirements in job advertisements[J]. EPJ Data Science,2019,8(1):1-20.

[11] 王东波,胡昊天,周鑫,等.基于深度学习的数据科学招聘实体自动抽取及分析研究[J].图书情报工作,2018,62(13):64-73.
WANG Dongbo,HU Haotian,ZHOU Xin,et al. Research of automatic extraction of entities of data science hiring with deep learning[J]. Library and Information Service,2018,62(13):64-73.

[12] 易新河,杨鹏,文益民.中文招聘文档中专业技能词抽取的跨域迁移学习[J].数据分析与知识发现,2022,6(2/3):274-288.
YI Xinhe,YANG Peng,WEN Yimin. Cross-domain transfer learning for recognizing professional skills from Chinese job postings[J]. Data Analysis and Knowledge Discovery,2022,6(2/3):274-288.

[13] LI X N,YAN H,QIU X P,et al. FLAT:Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:6836-6842.

[14] WU S,SONG X N,FENG Z H. MECT:Multi-metadata embedding based cross-transformer for Chinese named entity recognition[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021:1529-1539.

[15] LIU Y,LU J H,YANG J,et al. Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax[J]. Mathematical Biosciences and Engineering,2020,17(6):7819-7837.

[16] ZHU Y,WANG G. CAN-NER:Convolutional attention network for Chinese named entity recognition[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg,PA:Association for Computational Linguistics. 2019:3384-3393.

[17] 蒋翔,马建霞,袁慧. 基于 BiLSTM-IDCNN-CRF 模型的生态治理技术领域命名实体识别[J]. 计算机应用与软件,2021,38(3):134-141.
JIANG Xiang,MA Jianxia,YUAN Hui. Named entity recognition in the field of ecological management technology based on BiLSTM-IDCNN-CRF Model[J]. Computer Applications and Software,2021,38(3):134-141.

[18] LAFFERTY J,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,2001:282-289.

[19] HAMMERTON J. Named entity recognition with long short-term memory[C]//Proceedings of the Seventh Conference on Natural language Learning at HLT-NAACL. 2003:172-175.

[20] 黄梦醒,李梦龙,韩惠蕊. 基于电子病历的实体识别和知识图谱构建的研究[J]. 计算机应用研究,2019,36(12):3735-3739.
HUANG Mengxing,LI Menglong,HAN Huirui. Research on entity recognition and knowledge graph construction based on electronic medical records[J]. Application Research of Computers,2019,36(12):3735-3739.

[21] 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 1):15-19.
GUO Juncheng,WAN Gang,HU Xinjie,et al. Chinese resume named entity recognition based on BERT[J]. Journal of Computer Applications,2021,41(S1):15-19.

[22] ZHANG Y,YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg. PA:Association for Computational Linguistics. 2018:1554-1564.

[23] GUI T,MA R,ZHANG Q,et al. CNN-based Chinese NER with lexicon rethinking[C]//28th International Joint Conference on Artificial Intelligence. 2019:4982-4988.

[24] 赵卓,田侃,张殊,等. 基于预训练模型的文博数据命名实体识别方法[J]. 计算机应用,2022,42(增 1):48-53.
ZHAO Zhuo,TIAN Kan,ZHANG Shu,et al. Named entity recognition method for culture and museum data based on pretraining model[J]. Journal of Computer Applications,2022,42(S1):48-53.

(责任编辑:傅 游)