文章编号:1672-3961(2024)01-0091-09 DOI:10.6040/j.issn.1672-3961.0.2022.350

融合词汇信息与 GlobalPointer 的实体识别

李明键[1],李卫军[1,2*],王海荣[1,2]
(1.北方民族大学计算机科学与工程学院,宁夏 银川 750021;2.北方民族大学图形图像智能处理国家民委重点实验室,宁夏 银川 750021)

摘要:为了提升 GlobaiPointer 方法的实体边界区分性能,提出一种融合词汇信息与 GlobalPointer 的实体识别方法。对 SoftLexicon提取的词汇特征与字符相结合,采用 BiLSTM 网络与 RoPE 编码捕捉时序与相对位置信息构建全面特征,通过实体矩阵实现实体识别。对多个数据集进行试验,本研究提出的模型相较于其他基线模型,精确率、召回率、F_1 均有一定的提升,Weibo 数据集中 F_1 达到 71.33%、CMeEE 数据集中 F_1 达到 63.45%,表明本研究提出的模型架构能够进一步扩充语义表征,增强识别性能。

关键词:相对位置编码;词汇信息;实体识别;特征融合;神经网络
中图分类号:TP39 **文献标志码**:A
引用格式:李明键,李卫军,王海荣. 融合词汇信息与 GlobalPointer 的实体识别[J].山东大学学报(工学版),2024,54(1):91-99.
LI Mingjian, LI Weijun, WANG Hairong. Entity recognition based on lexicon information and GlobalPointer[J]. Journal of Shandong University (Engineering Science), 2024, 54(1):91-99.

Entity recognition based on lexicon information and GlobalPointer

LI Mingjian[1], LI Weijun[1,2*], WANG Hairong[1,2]
(1. School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, Ningxia, China; 2. The Key Laboratory of Images & Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, Ningxia, China)

Abstract: In order to improve the entity boundary differentiation performance of GlobalPointer, an entity recognition method integrating lexicon information and globalpointer was proposed to enhance the recognition performance. For characters, softlexcion was used to extract vocabulary features and combine them with characters. BiLSTM network and RoPE code were used to capture timing and relative position information to construct comprehensive features. Entity recognition was realized through entity matrix. Experiments were carried out on multiple datasets. Compared with other baseline models, the model had made some progress in the metrics of precision, recall and F_1. The F_1 in Weibo dataset had reached 71.33%, and the F_1 in CMeEE dataset had reached 63.45%. It indicated that the model architecture could further expand semantic representation and enhance recognition performance.
Keywords: relative position code; lexicon information; entity identification; features fusion; neural network

0 引言

命名实体识别(name entity recognition, NER)在信息抽取、自然语言处理等领域有着重要地位,其含义是通过对输入文本序列进行特征标注从而发现专有词语,例如人名、地名、机构名等,这类专有词语随着所研究的数据集不同而改变。目前命名实体识别技术中针对扁平实体结构主要采用序列标注方式,而对嵌套实体结构主要采用基于跨度方法,通过识别实体头尾,确定实体位置。GlobalPointer 模型采用实体矩阵方法能够良好识别扁平

收稿日期:2022-10-14
基金项目:宁夏自然科学基金资助项目(2021AAC03215);北方民族大学重点科研项目(2021JCYJ12)
第一作者简介:李明键(1997—),男,四川江油人,硕士研究生,主要研究方向为实体识别、机器学习。E-mail:1143311329@qq.com
*通信作者简介:李卫军(1979—),男,陕西渭南人,讲师,硕士生导师,博士,主要研究方向为本体的构建与重用、知识图谱的构建。
E-mail:lwj@nmu.edu.cn

与嵌套实体结构,但 GlobalPointer 中未能充分开发字符特征,影响了实体边界区分,限制了识别性能。为此,本研究设计了融合词汇信息与 GlobalPointer 的实体识别模型。

目前命名实体识别的方法主要有 3 种。

(1)基于规则、词典匹配的方法。通过专家手工构建实体组成规则模板以及建立专有词汇实体词典,以文字序列匹配规则与词典的方式识别专有词汇。但该方法移植性低、复杂度高。

(2)基于机器学习的方法。主要使用条件随机场[1](conditional random field, CRF)、隐马尔可夫模型[2](hidden markov model, HMM)、支持向量机[3](support vector machine, SVM)等方法。由于 CRF 特征捕获全面,能更好地利用上下文信息,具有很强的推理能力,因此相比于 HMM、SVM 在实体识别中应用更广。

(3)基于深度学习的方法。通过深层神经网络逐层提取数据的特征并使用非线性激活函数对特征进行学习、整合以获取数据的潜藏信息。当前主流方式是使用 BiLSTM-CRF[4](bi-directional long short-term memory, BiLSTM)模型以及基于卷积神经网络[5](convolutional neural networks, CNN)模型进行实体识别,在 MSRA 数据、人民日报数据等数据集中取得了较好的效果。随着注意力机制的发展特别是 Transformer[6] 的提出为实体识别注入了新动力,由此一些学者将循环神经网络(recurrent neural network, RNN)与注意力机制相结合,利用注意力机制提取特征之间的关联度使得特征结合更完善从而提升识别效果。文献[7]提出一种结合自注意力的 BiLSTM-CRF 的电子病历命名实体识别方法,弥补现有方法不能很好捕获电子病历实体之间的长距离依赖关系的缺陷,在 CCKS 数据集上与现有方法相比大大提升了效果;文献[8]提出一种基于 Transformer 编码器的中文命名实体识别方法,在模型的字嵌入层中使用结合词典的字向量编码方法,从而让字向量包含了词语信息并改进了 Transformer 的注意力计算方式;文献[9]提出一种基于多头自注意力神经网络的中文临床命名实体识别方法,使用一种新颖的融合领域词典的字符级特征表示方法,结合多头自注意力机制和 BiLSTM-CRF 准确地捕获字符间潜在的依赖权重、语境和语义关联等多方面的特征。

在上述方法基础上为提升识别效果有学者还从以下方面进行了研究。

(1)结合预训练模型。预训练模型是通过大量数据训练得来,其参数更具有实际意义,因此从词向量的角度出发,利用预训练模型获得数据的普遍特征使得模型训练更具有泛化性以及加速收敛。文献[10]提出一种基于 BERT(bidirectional encoder representations from transformers)的中文电子简历命名实体识别方法,该方法解决了建立数据中实体提取方法效率低、迁移能力差的问题;文献[11]提出一种基于 BERT 的中文简历命名实体识别方法,将 BERT 与 BiLSTM 相结合充分发掘中文简历数据中所蕴含的信息,提高构建社交网络知识图谱和档案知识图谱的实体丰富度。

(2)增加词汇特征。在字符特征中融入词汇特征能够加强字符表征,提升识别效果。文献[12]提出的 Lattice-LSTM 方法将字符与所匹配的词汇融合,增强了字符语义信息并且通过 BiLSTM 捕获时序特征在多种数据上都优于之前算法,但存在移植性差、信息损失等问题;文献[13]在此问题上提出 Soft-Lexicon 模型,以一种简单高效的方式融合字词,通过 BMES(begin medial end single) 4 种标志进行匹配查找每个字符对应的词汇,然后以词频作为特征融合中心进行词汇增强,试验结果表明相较于 Lattice-LSTM 性能有很大提升;文献[14]提出一种基于新词发现和 Lattice-LSTM 的中文医疗命名实体识别方法,将新词发现与实体识别模型结合,在医疗文本中取得了不错的效果。

通过上述研究路线可以知道,增加外部特征(如词汇、BERT 预训练向量)能够有效提升识别效果。但 GlobalPointer 缺乏词汇信息引入,同时未能完善上下文语义信息,因此本研究在此基础上提出一种融合词汇信息与 GlobalPointer 的模型,即 Lx-GlobalPointer。考虑到每个字符对应的词汇特征也具有时序性,本研究针对字符向量在 BERT 后采用 BiLSTM 进行时序建模,以有效利用预训练词向量;将词汇信息表示使用 BiLSTM 建模以更好地表征外部特征,将字符信息与外部特征通过相加的方式,以此完成嵌入特征的结合强化语义信息,从而完善特征表示。

1 Lx-GlobalPointer 模型

在字符特征基础上增加词汇以加强数据表征,使用 BiLSTM 以及相对位置编码对词嵌入向量建模深层语义信息,构建实体矩阵进行实体片段抽取以实现最终实体识别。总体架构分为词汇表示层、特征融合层、GlobalPointer 层,具体模型如图 1 所示。

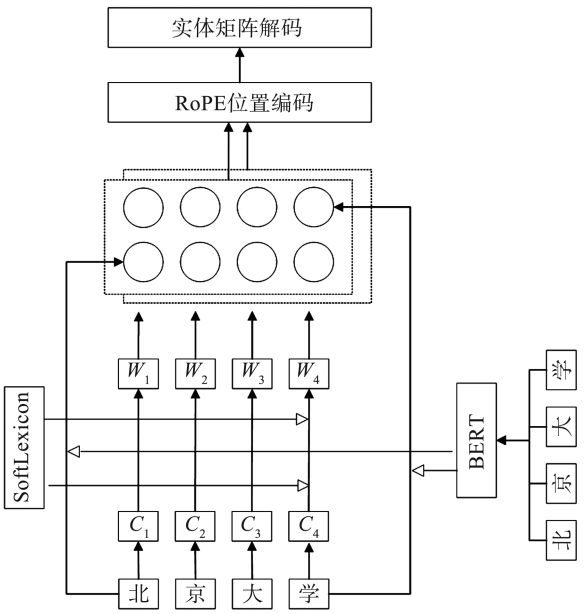

图 1　Lx-GlobalPointer 模型图
Fig.1　Lx-GlobalPointer model

1.1　词汇表示层

Softlexicon 通过 B、M、E、S 这 4 种结构引入词汇,解决了词汇损失的问题,具体实现如图 2 所示。

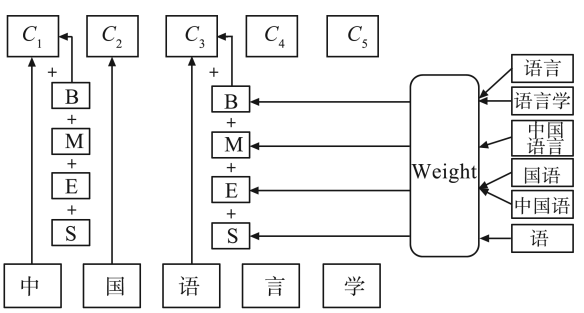

图 2　SoftLexicon 模型图
Fig.2　SoftLexicon model

以语字为例,B 表示匹配以该字符为首位的词语如语言、语言学,M 表示匹配该字符处于中间的词语如中国语言,E 表示匹配该字符处于结尾的词语如国语、中国语,S 表示该字符本身。通过该方法收集每个字符所对应的词汇,获取到词汇的完整信息,接下来通过词频、拼接方式表征词汇信息,B、M、E、S 结构的构建公式为:

$$\boldsymbol{B}(c_i)=\{w_{i,k},\forall w_{i,k}\in L,i<k\leqslant n\},\tag{1}$$

$$\boldsymbol{M}(c_i)=\{w_{j,k},\forall w_{j,k}\in L,1\leqslant j<i<k\leqslant n\},\tag{2}$$

$$\boldsymbol{E}(c_i)=\{w_{j,i},\forall w_{j,i}\in L,1\leqslant j<i\},\tag{3}$$

$$\boldsymbol{S}(c_i)=\{c_i,\exists c_i\in L\},\tag{4}$$

$$v^s(s)=\frac{4}{Z}\sum_{W\in S}z(w)\boldsymbol{e}^w(w),\tag{5}$$

$$Z=\sum_{\boldsymbol{w}\in\boldsymbol{B}\cup\boldsymbol{M}\cup\boldsymbol{E}\cup\boldsymbol{S}}z(\boldsymbol{w}),\tag{6}$$

$$\boldsymbol{e}^s(\boldsymbol{B},\boldsymbol{M},\boldsymbol{E},\boldsymbol{S})=[v^s(\boldsymbol{B});v^s(\boldsymbol{M});v^s(\boldsymbol{E});v^s(\boldsymbol{S})],\tag{7}$$

式中,$\boldsymbol{B}(c_i)$、$\boldsymbol{M}(c_i)$、$\boldsymbol{E}(c_i)$、$\boldsymbol{S}(c_i)$分别是字符 c_i 匹配到对应结构的词汇向量集,L 表示词典,$w_{i,k}$、$w_{j,k}$、$w_{j,i}$所对应结构的词汇,c_i 表示单个字符,i、j、k 表示索引位置,n 为文本序列,$v^s(s)$表示对单个结构的词汇信息进行归一整合,$z(w)$为单个词汇的词频,$\boldsymbol{e}^w(w)$为词汇嵌入矩阵,Z 为 4 种词汇结构的词频之和,$\boldsymbol{e}^s(\boldsymbol{B},\boldsymbol{M},\boldsymbol{E},\boldsymbol{S})$表示将 4 种结构的词汇向量拼接得到最终特征表示。

1.2　特征融合层

BiLSTM 由 2 个 LSTM 组建而成[15],以获取全局特征。单个 LSTM 结构如图 3 所示。

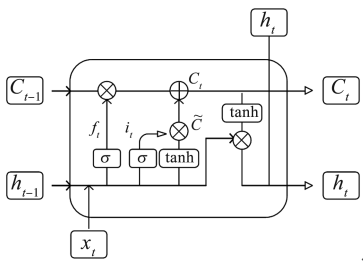

图 3　LSTM 结构图
Fig.3　LSTM model

图 3 中,f_t、i_t、o_t、h_t 分别为遗忘门、记忆门、输出门、隐层输出,σ 为 sigmoid 激活函数,tanh 为双曲正切激活函数。在 t 时刻的输入为 x,其输入 h_t 是通过遗忘门、记忆门、输出门的计算所得,公式表示为:

$$f_t=\sigma(\boldsymbol{W}_f\cdot[h_{t-1},x_t]+b_f),\tag{8}$$

$$i_t=\sigma(\boldsymbol{W}_i\cdot[h_{t-1},x_t]+b_i),\tag{9}$$

$$\widetilde{C}=\tanh(\boldsymbol{W}_c\cdot[h_{t-1},x_t]+b_c),\tag{10}$$

$$C_t=f_t\odot C_{t-1}+i_t\odot\widetilde{C}_t,\tag{11}$$

$$o_t=\sigma(\boldsymbol{W}_o\cdot[h_{t-1},x_t]+b_o),\tag{12}$$

$$h_t=o_t\odot\tanh(C_t),\tag{13}$$

式中,$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_c$、$\boldsymbol{W}_o$ 为矩阵参数,b_f、b_i、b_c、b_o 为偏置参数,$\widetilde{C}$、C_t 分别为临时细胞状态与此时细胞状态。

BiLSTM 使用正向和反向的 LSTM 捕捉过去和将来的信息,拼接构成双向结果,下一步采用 BiLSTM 网络对词汇与字符特征进行优化,公式表示为:

$$\text{BiLSTM\_output\_char}\leftarrow\text{BiLSTM}(\text{BERT}(x^c)),\tag{14}$$

$$\boldsymbol{x}^w\leftarrow\boldsymbol{e}^s(\boldsymbol{B},\boldsymbol{M},\boldsymbol{E},\boldsymbol{S}),\tag{15}$$

$$\text{BiLSTM\_output\_char}\leftarrow\text{BiLSTM}(\boldsymbol{x}^w)+\text{BiLSTM\_output\_char},\tag{16}$$

式中:BiLSTM_output_char 表示使用 BERT[16] 预训练模型对字符 x^c 进行向量初始化并由 BiLSTM 网络得到字符信息的广义语义表示;$\boldsymbol{x}^w$ 表示通过 SoftLexcicon 结构得到字符的词汇特征表示;BiLSTM_

output 表示将词汇特征表示输入到 BiLSTM 层进行序列建模并与字符 BiLSTM 结果相加得到最终时序输出,以此完成两者信息的融合。

本研究与 GlobalPointer 的预训练模型均采用 BERT。BERT 是使用 MLM(masked language model)和双向 Transformer 进行训练并构建的模型,相比于传统 word2vec[17] 和 Glove[18] 能够生成结合上下文信息的深层双向语言表征,有效解决字的多义性问题。

1.3 GlobalPointer 层

GlobalPointer 是文献[19]提出的一种解决实体识别的统一方法,能够有效处理嵌套与非嵌套实体识别问题,其识别方式与 TPLinker[20] 相近,主要思路是通过构建实体矩阵实现模型的训练与预测。GlobalPointer 总体由 3 部分组成:核心计算公式、相对位置编码、损失函数,结构如图 4 所示。

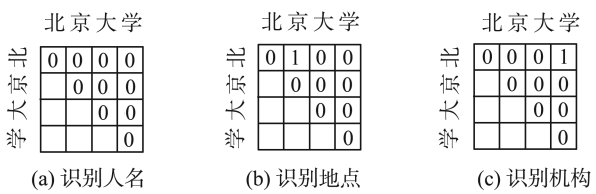

图 4 GlobalPointer 结构图
Fig.4 GlobalPointer structure

1.3.1 核心计算公式

如图 4 所示,GlobalPointer 是通过构建实体矩阵完成每个子序列的类别区分,之后通过首尾标记、片段分类的方法实现实体识别。其中横向代表开始位置,纵向代表结束位置,例如“北京”表示地点,即在[0,1]位置上标记 1,“北京大学”代表机构,即在[0,3]位置上标记 1。公式表示为:

$$s_a(i,j)=\boldsymbol{q}_{i,a}^{\mathrm{T}},\boldsymbol{k}_{j,a}, \tag{17}$$

$$\boldsymbol{q}_{i,a}=\boldsymbol{w}_{q,a}\boldsymbol{h}_i+b_{q,a}, \tag{18}$$

$$\boldsymbol{k}_{j,a}=\boldsymbol{w}_{k,a}\boldsymbol{h}_j+b_{k,a}, \tag{19}$$

式中:S_a 为候选片段 $[i,j]$ 表示的实体类型 a 的得分,i、j 分别为片段的开始、结束位置;$\boldsymbol{h}_i$ 为字符向量;$\boldsymbol{w}_{q,a}$、$\boldsymbol{w}_{k,a}$、$b_{q,a}$、$b_{k,a}$ 分别为矩阵参数、偏置参数;$\boldsymbol{q}_{i,a}$,$\boldsymbol{k}_{j,a}$ 为对输入序列经过神经网络编码所得的输出向量。

1.3.2 相对位置编码

观察公式(17)可以得知,实体分数计算中字符向量的表示特征单一、包含的语义信息少、缺乏位置信息,因此 GlobalPointer 使用 RoPE[21] 进行位置编码,公式表示为:

$$s_a(i,j)=(\boldsymbol{R}_i\boldsymbol{q}_{i,a})^{\mathrm{T}}(\boldsymbol{R}_j\boldsymbol{k}_{j,a})=\boldsymbol{q}_{i,a}^{\mathrm{T}}\boldsymbol{R}_i^{\mathrm{T}}\boldsymbol{R}_j\boldsymbol{k}_{j,a}=\boldsymbol{q}_{i,a}^{\mathrm{T}}\boldsymbol{R}_{i-j}\boldsymbol{k}_{j,a}, \tag{20}$$

$$s_a(i,j)=(\boldsymbol{R}_i\mathrm{BiLSTM\_output}_{i,a})^{\mathrm{T}}(\boldsymbol{R}_j\mathrm{BiLSTM\_output}_{j,a}), \tag{21}$$

式中:i、j 为字符的位置索引;$\boldsymbol{R}_i$、$\boldsymbol{R}_j$ 为正交矩阵;$\boldsymbol{R}_{i-j}$ 为 $\boldsymbol{R}_i^{\mathrm{T}}$ 与 $\boldsymbol{R}_j$ 的内积,表示相对位置向量。之后将特征融合层输出带入 $\boldsymbol{q}_{i,a}$ $\boldsymbol{k}_{j,a}$,以此实现 GlobalPointer 与词汇、字符的融合。

1.3.3 损失函数

GlobalPointer 的损失函数是比较目标类别与非目标类别的得分,以此平衡每一类别的权重,解决类别不均衡问题[22]。公式表示为:

$$\log\Big(1+\sum_{(i,j)\in P_a}\mathrm{e}^{-s_a(i,j)}\Big)+\log\Big(1+\sum_{(i,j)\in Q_a}\mathrm{e}^{s_a(i,j)}\Big), \tag{22}$$

式中,P_a 是该样本的所有类型为 a 的实体的首尾集合,Q_a 是该样本的所有非实体或者类型非 a 的实体的首尾集合,log 为 logsumexp 函数。

2 试验

2.1 试验数据集

本研究使用的公开数据集有 Weibo[23] 数据集和 CMeEE[24] 数据集,具体统计如表 1 所示。

表 1 数据统计
Table 1 Data statistics 单位:个

数据集	统计数量					
	训练		验证		测试	
	句子	字符	句子	字符	句子	字符
Weibo	1 400	73 800	270	14 800	270	14 800
CMeEE	15 000	812 300	5 000	270 400	3 000	165 700

Weibo 数据集有 8 种实体,分别是 PER.NOM、LOC.NAM、PER.NAM、GPE.NAM、ORG.NAM、ORG.NOM、LOC.NOM、GRE.NOM,其中 PER 表示人名,LOC 表示地名,GPE 表示地缘政治实体,ORG 表示组织机构实体,NAM 表示特指,NOM 表示泛指;CMeEE 数据集有 9 种实体并且含有嵌套实体,分别为 pro、dis、sym、ite、bod、dru、mic、equ、dep,分别表示医疗程序、疾病、临床表现、医学检验项目、身体、药物、微生物类、医疗设备、科室实体。

2.2 对比模型

为了验证本研究方法的有效性,选择了近来相关主流模型进行验证,分别是:SoftLexicon、Flat-lattice[25]、Lattice-LSTM[12]、BERT-BiLSTM-self-attention-CRF[26]、Transformer-HMM-BERT-NER[27]、GlobalPointer、WSA-CNER-BERT[28]。

2.3 评价方法

本试验采用精确率 P、召回率 R、F_1(F-

measure)评价指标进行评价,计算公式为:

$$P=R_N/P_N\times100\%, \quad (23)$$

$$R=R_N/G_N\times100\%, \quad (24)$$

$$F_1=2PR/(P+R)\times100\%, \quad (25)$$

式中,R_N 为预测结果中正确的实体数,P_N 为预测的所有实体数,G_N 为测试数据中正确的实体数。

2.4　试验设置

本研究的模型参数设置:pytorch 版本 1.8.1、Python版本 3.6.5、未使用预训练词嵌入(gigaword_chn.all.a2b.uni.ite50.vec)、Adam 优化器、迭代次数为 50、Batch 为 48,其余如表 2 所示。

表 2　参数设置

Table 2　Parameter setting

字符向量维度	隐藏层维度	词汇向量维度	droupout	BERT 层学习率	BiLSTM 层学习率
100	300	100	0.5	0.000 01	0.000 2

由于本研究使用了 BiLSTM 联合 BERT,因此采用分层学习率。本研究方法中所涉及的词汇是通过 jieba 词典对数据集分词所得,通过分词能够采集到字符的外部信息,使得实体边界区分度更高,具体示例如表 3 所示。

表 3　分词示例

Table 3　Example of participle

	示例
句子	科技全方位资讯智能,快捷的汽车生活需要有三屏一云爱你
分词结果	科技、全方位、方位、资讯、智能、快捷、汽车、生活、需要

2.5　试验分析

本节对比试验的结果均是对比模型原论文中的试验结果,由于 GlobalPointer 模型中没有 Weibo 数据集的结果,因此使用原方法参数重新试验。具体结果如表 4 所示。

表 4　Weibo-F_1 试验结果

Table 4　Weibo-F_1 experimental results

对比模型	F_1/%		
	NE	NM	Overall
Lx-GlobalPointer	69.81	**67.88**	**71.33**
SoftLexicon	70.94	67.02	70.50
Lattice-LSTM	53.04	62.25	58.79
Flat-lattice			68.55
GlobalPointer	67.99	67.40	70.04
BERT-BiLSTM-self-attention-CRF			69.00
Transformer-HMM-BERT-NER			68.00
WSA-CNER-BERT	**74.07**	67.82	71.04

注:NE、NM、Overall 分别表示特指实体、代指实体、两者综合。

由表 4 可以看出在 Weibo 数据集中,与 SoftLexicon、Flat-Lattice、BERT-BiLSTM-self-attention-CRF、Transformer-HMM-Bert-NER、GlobalPointer、WSA-CNER-BERT 相比,Lx-GlobalPointer 在 Weibo-Overall 上 F_1 分别提高 0.83%、2.78%、2.33%、3.33%、1.29%、0.29%;与 SoftLexicon、GlobalPointer、WSA-CNER-BERT 相比 Lx-GlobalPointer 在 Weibo-NE 上的 F_1 分别下降 1.13%、提升 1.82%、下降 4.26%;与 SoftLexicon、GlobalPointer、WSA-CNER-BERT 相比 Lx-GlobalPointer 在 Weibo-NM 上 F_1 分别提高 0.86%、0.48%、0.06%;F_1 的提高表明了 Lx-GlobalPointer 方法是有效果的,其原因是 Weibo 数据由于数据格式不固定、相似度低导致词汇的差异性大,而引入词汇信息能够进一步完善 Weibo 数据集的字符底层特征,提升识别效果。

2.6　消融试验

为验证本研究提出方法的有效性,针对 Weibo-Overall 数据集还做了消融试验如下。

Lx-GlobalPointer(w/o Lexicon):在本研究结构上去掉词汇信息,替换为字符信息。

SoftLexicon-Lx:按照本研究网络框架对词汇进行融合,字符、词汇的特征仍然使用 SoftLexicon 原论文的方法。

表 5 中前 3 种模型为词汇特征验证,后 2 种模型为网络框架通用性验证,因此试验基于表 2 参数,这样更能体现出公平性、一致性。具体结果如表 5 所示。

表 5　Weibo-F_1 消融试验结果

Table 5　Results of Weibo-F_1 ablation test

单位:%

对比模型	P	R	F_1
Lx-GlobalPointer	74.02	**68.84**	**71.33**
Lx-GlobalPointer (w/o Lexicon)	69.37	70.04	69.71
GlobalPointer	**76.87**	64.25	70.00
SoftLexicon-Lx	68.23	59.66	63.66
SoftLexicon	66.58	60.63	63.46

从表 5 可以看出,相比于 GlobalPointer、Lx-GlobalPointer(w/o Lexicon),本研究所提出的 Lx-GlobalPointer F_1 提升了 1.33%、1.62%;SoftLexicon-Lx 相比于 SoftLexicon F_1 提升了 0.20%。由此可以得知:(1)Lx-GlobalPointer 方法对比于其他方法 F_1 均有一定的提升,说明本研究从词汇特征与网络的构建出发所提的改进方法是有效果的;(2)Lx-GlobalPointer 与 Lx-GlobalPointer(w/o Lexicon)的

结果可以看出相同网络条件下，加入词汇信息能够有效完善网络输入特征，进一步提升性能；（3）联合SoftLexicon与SoftLexicon-Lx的结果，可以得知，同样是引入词汇信息，本研究框架相比于SoftLexicon方法有更优的效果，验证了网络框架的通用性，表明本研究使用BiLSTM对词汇进行序列建模并采用相加方式以完成词汇与字符的融合的方法是有效果的。

2.7 嵌套实体识别分析

为验证本研究方法在嵌套实体中的应用，在CMeEE数据集中做了试验，数据集与试验结果来自中文医疗信息处理挑战榜 https://tianchi.aliyun.com/cblue，对比方法分别是Lx-GlobalPointer、GlobalPointer、SoftLexicon。试验设置：迭代次数为10，Batch为8，未使用预训练词嵌入文件，学习率设置为0.000 01，其余参考表2，试验结果如表6所示。

表6 CMeEE试验结果

Table 6 CMeEE experimental results 单位：%

对比模型	P	R	F_1
Lx-GlobalPointer	**67.72**	**59.69**	**63.45**
GlobalPointer	65.85	58.81	62.13
SoftLexicon	65.50	54.72	59.62

从表6可以看出，Lx-GlobalPointer相比GlobalPointer F_1提升了1.32%；相比SoftLexicon F_1提升了3.83%；GlobalPointer相比SoftLexicon F_1提升了2.51%。由此可以得出如下结论。

（1）SoftLexicon采用CRF方式进行解码并不能很好地解决嵌套实体，而GlobalPointer方法将实体类型分开预测，不再依赖总体线性标签得分，从而能够有效处理嵌套实体。

（2）相比SoftLexicon与GlobalPointer，Lx-GlobalPointer能够有效提高P、R、F_1，表明引入词汇丰富了底部数据特征，双层BiLSTM的使用加深了字词向量的提取特征，从而加强了字符的区分性，使得识别效果显著提升。

（3）GlobalPointer在本地Pytorch框架上运行，F_1是62.13%，GlobalPointer原方法使用keras在CMeEE上的F_1是65.98%，二者的差距为3.85%。在RoBerta-large模型中试验，试验设置：迭代次数为10，Batch为4，未使用预训练词嵌入文件，学习率设置为0.000 01，其余参考表2、GlobalPointer设置。本地GlobalPointer F_1是66.45%，原方法F_1是67%及以上，二者的差距为0.55%～1.55%。Lx-GlobalPointer上试验结果是67.01%，与GlobalPointer相比提升了0.56%。在相同试验设置下运行结果表明，本研究方法对性能提升有一定帮助。

2.8 迭代试验分析

为更深层次验证本研究方法，在Weibo-Overall、CMeEE数据集上对训练过程中每次迭代的验证集的F_1与迭代次数进行记录，观测训练中模型的性能变化。试验设置如下。

Lx-GlobalPointer：Weibo-Overall数据集、CMeEE数据集中参数设置与表3与2.7节一致。

GlobalPointer：参数设置与Lx-GlobalPointer一致。

SoftLexicon：由2.7节可知，SoftLexicon在嵌套数据中性能欠佳，因此主要针对Weibo-Overall数据集进行试验，其中关键参数与表2一致，其余参照原论文设置。试验结果如图5、6及表7、8所示。

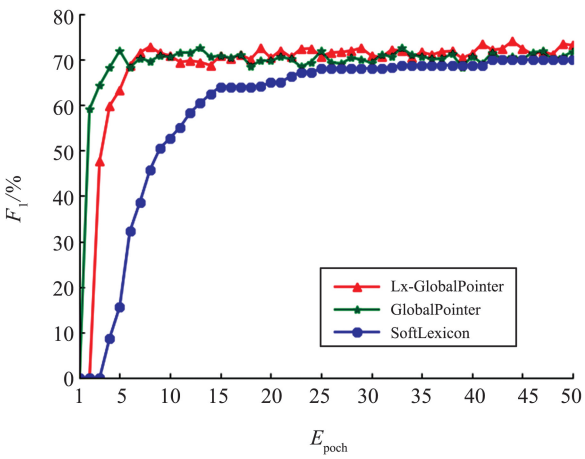

图5 Weibo-Overall验证集F_1分析图

Fig.5 The F_1 of verification set change with epoch in Weibo-Overall

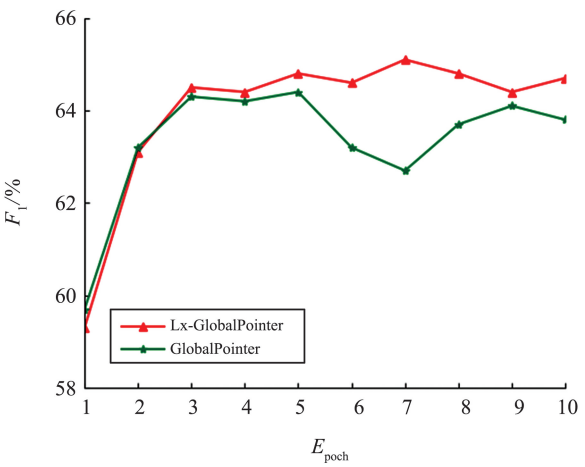

图6 CMeEE验证集F_1分析图

Fig.6 The F_1 of verification set change with epoch in CMeEE

表 7　Weibo-Overall 验证集实体数量分析
Table 7　The number statistic of entities for verification set in Weibo-Overall　　单位:个

实体类型	Lx-GlobalPointer 预测实体类型数量	GlobalPointer 预测实体类型数量	SoftLexicon 预测实体类型数量
PER.NOM	**170**	169	169
LOC.NAM	**5**	1	0
PER.NAM	68	61	**72**
GPE.NAM	**23**	21	17
ORG.NAM	**18**	14	16
ORG.NAM	**4**	3	0
ORG.NAM	**3**	**3**	0
ORG.NAM	0	0	0

表 8　CMeEE 验证集实体数量分析
Table 8　The number statistic of entities for verification set in CMeEE　　单位:个

实体类型	Lx-GlobalPointer 预测实体类型数量	GlobalPointer 预测实体类型数量
pro	1 266	**1 322**
dis	3 681	**3 992**
sym	**1 848**	1 362
ite	**394**	288
bod	**4 061**	4 025
dru	**1 200**	1 134
mic	**489**	450
equ	**124**	99
dep	42	**55**

由图 5、6、表 7、8 可知。

(1) 图 5 为 Weibo 验证数据集 F_1 在不同迭代下的变化趋势。SoftLexicon 方法 F_1 一直处于较低水平;Lx-GlobalPointer 从第 5 轮迭代 F_1 开始超越 GlobalPointer,然后总体上一直维持该优势;最终 3 种模型从第 32 轮迭代 F_1 开始趋于稳定。GlobalPointer、Lx-GlobalPointer、SoftLexicon F_1 的最高值 72.52%、73.96%、69.90%,分别是在第 12 轮、第 43 轮、第 41 轮迭代取得。

(2) 图 6 为 CMeEE 验证数据集 F_1 在不同迭代下的变化趋势。Lx-GlobalPointer 从第 2 轮迭代 F_1 开始超越 GlobalPointer,到结束一直保持领先优势;Lx-GlobalPointer、GlobalPointer 的最高 F_1 为 65.05%、64.43%,分别是在第 6 轮、第 4 轮迭代取得。此外 GlobalPointer 相比于 Lx-GlobalPointer 的曲线变化幅度更大,稳定性相比较低。

(3) 表 7 为 Weibo 数据集中各个实体类型的数量统计。在 8 种实体类型中,Lx-GlobalPointer 预测数量有 6 种实体高于 GlobalPointer,有 6 种实体高于 SoftLexicon;GlobalPointer 预测数量有 4 种实体高于 SoftLexicon;SoftLexicon 预测数量有 1 种实体高于 Lx-GlobalPointer; SoftLexicon 预测数量有 2 种实体高于 GlobalPointer。在 PER.NAM 类型中,Lx-GlobalPointer 效果低于 SoftLexicon,这是因为该类实体主要为特指人名信息,这类实体中蕴含的词汇信息缺乏,词汇特征加入会引起干扰,影响识别效果;而在 LOC.NAM 类型中,实体为特指地名,例如:云台山、沧海路等这类实体,其中词汇信息丰富,可以提取出沧海、台山等词汇,以加强语义边界提升识别效果。

(4) 表 8 为 CMeEE 数据集中各个实体类型的数量统计。在 9 种实体类型中,Lx-GlobalPointer 预测数量有 6 种实体高于 GlobalPointer;GlobalPointer 预测数量有 3 种实体高于 Lx-GlobalPointer。在 pro、dis、dep 类型中,Lx-GlobalPointer 效果低于 GlobalPointer,同上,因为实体的词汇信息不充分、黏合性低,影响了识别效果。

上述分析可以得知,GlobalPointer 的综合性能优于 SoftLexicon,而在此基础上构建的 Lx-GlobalPointer 拥有更高的稳定性、鲁棒性,对特征的抽取也更加全面、深入,证明了本研究中词汇、模型框架使用是切实有效的。

2.9　实例分析

针对嵌套与非嵌套实体,通过具体例子证明本研究方法的有效性,具体结果如下。

(1)非嵌套实体例子

“每个孩子都是一个天使,折翼来到人间,每个母亲都会守护着自己的小天使”,正确实体为“孩子”、“天使”、“母亲”、“小天使”。结果表明:Lx-GlobalPointer 模型能够正确识别出所有的实体,GlobalPointer 模型除了“天使”其余实体均能识别,SoftLexicon 模型仅识别出了“孩子”和“母亲”实体。

(2)嵌套实体例子

“胸廓逐渐出现上窄下宽的圆锥形形态”,正确实体为“胸廓逐渐出现上窄下宽的圆锥形形态”、“胸廓”。结果表明:Lx-GlobalPointer 模型与 GlobalPointer 模型能够正确识别出所有的实体;SoftLexicon 模型仅识别出了外层实体“胸廓逐渐出现上窄下宽的圆锥形形态”,未能将内部实体“胸廓”识别出来。

综上所述,可以知道:

(1) SoftLexicon 对“小天使”的 CRF 解码标签为“B-PER.NOM、B-PER.NOM、E-PER.NOM”,而正确的解码标签应该是“B-PER.NOM、M-PER.NOM、

E-PER.NOM”，可以看出由于错误标签的得分高于正确标签，从而导致了实体的误判。

（2）相比于 GlobalPointer，Lx-GlobalPointer 能够识别出更多正确实体，表明了本研究引入词汇特征、BiLSTM 网络，从增强字符基础特征出发能够提升识别效果，验证了本研究方法的有效性。

3 结语

本研究提出了融合词汇信息与 GlobalPointer 的识别方法。该模型能够获取外部信息，使向量表征更加合理、丰富，同时进行相对位置编码以及时序特征建模以增强数据联系，学习到深层语义信息。试验证明，相比于其他基线模型，本研究方法有着更优异的性能，表明加入词汇特征以及丰富神经网络结构，从特征构建的广度、深度切入，对命名实体识别模型性能提升都有帮助。

但本研究仍存在一些不足：仅引入词汇特征未能完全发掘外部信息的作用；实体矩阵间的联系相对减弱，无法有效利用类别转移关系。下一步工作将采取有效措施扩展外部特征以及利用类别转移，提高识别效果。

参考文献：

[1] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]//Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. [S.l.]: ACM, 2003: 188-191.

[2] BIKEL D M, MILLER S, SCHWARTZ R, et al. Nymble: a high-performance learning name-finder[EB/OL]. (1998-03-27)[2021-09-15]. https://arxiv.org/pdf/cmp-lg/9803003.

[3] JU Z, WANG J, ZHU F. Named entity recognition from biomedical text using SVM[C]//2011 5th International Conference on Bioinformatics and Biomedical Engineering. New York, USA: IEEE, 2011: 1-4.

[4] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-08-09)[2021-07-24]. https://arxiv.org/pdf/1508.01991.

[5] DONG X, QIAN L, GUAN Y, et al. A multiclass classification method based on deep learning for named entity recognition in electronic medical records[C]//2016 New York: Scientific Data Summit (NYSDS). New York, USA: IEEE, 2016: 1-10.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2021-03-02]. https://arxiv.org/pdf/1706.03762.

[7] 曾青霞，熊旺平，杜建强，等. 结合自注意力的 BiLSTM-CRF 的电子病历命名实体识别[J]. 计算机应用与软件，2021，38(3)：159-162.
ZENG Qingxia, XIONG Wangping, DU Jianqiang, et al. Naming entity recognition of electronic medical records based on self-attention BiLSTM-CRF[J]. Computer Application and Software, 2021, 38(3): 159-162.

[8] 司逸晨，管有庆. 基于 Transformer 编码器的中文命名实体识别模型[J]. 计算机工程，2022，48(7)：66-72.
SI Yichen, GUAN Youqing. Chinese named entity recognition model based on transformer encoder[J]. Computer Engineering, 2022, 48(7): 66-72.

[9] 罗熹，夏先运，安莹，等. 结合多头自注意力机制与 BiLSTM-CRF 的中文临床实体识别[J]. 湖南大学学报（自然科学版），2021，48(4)：45-55.
LUO Xi, XIA Xianyun, AN Ying, et al. Chinese clinical entity recognition combined with multi-head self-attention mechanism and BiLSTM-CRF[J]. Journal of Hunan University (Natural Science Edition), 2021, 48(4): 45-55.

[10] 王传涛，丁林楷，杨学鑫，等. 基于 BERT 的中文电子简历命名实体识别[J]. 中国科技论文，2021，16(7)：770-775.
WANG Chuantao, DING Linkai, YANG Xuexin, et al. Chinese electronic resume named entity recognition based on BERT[J]. Chinese Science and Technology Paper, 2021, 16(7): 770-775.

[11] 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 Juncheng, WAN Gang, HU Xinjie, et al. Chinese resume named entity recognition based on BERT[J]. Computer Application, 2021, 41(Suppl.1): 15-19.

[12] ZHANG Y, YANG J. Chinese NER using lattice LSTM[EB/OL]. (2018-05-05)[2021-05-06]. https://arxiv.org/pdf/1805.02023.

[13] MA R, PENG M, ZHANG Q, et al. Simplify the usage of lexicon in chinese NER[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, USA: ACL, 2020: 5951-5960.

[14] 赵耀全，车超，张强. 基于新词发现和 Lattice-LSTM 的中文医疗命名实体识别[J]. 计算机应用与软件，2021，38(1)：161-165.
ZHAO Yaoquan, CHE Chao, ZHANG Qiang. Chinese medical named entity recognition based on neologism discovery and Lattice-LSTM [J]. Computer Application and Software, 2021, 38(1): 161-165.

[15] GRAVES A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2013: 6645-6649.

[16] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11) [2021-05-03]. https://arxiv.org/pdf/1810.04805.

[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-01-16) [2021-12-18]. http://arxiv. org/abs/1301.3781.

[18] 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. Stroudsburg, USA: ACL, 2014: 1532.

[19] SU J L, MURTADHA A, PAN S, et al. GlobalPointer: novel efficient span-based approach for named entity recognition [EB/OL]. (2022-08-15) [2022-09-12]. https://arxiv.org/abs/2106.08087.

[20] WANG Y, YU B, ZHANG Y, et al. TPLinker: single-stage joint extraction of entities and relations through token pair linking[C]//Proceedings of the 28 th International Conference on Computational Linguistics. Barcelona, Spain: International Committee on Computational Linguistics, 2020: 1572-1582.

[21] SU J L, LU Y, PAN S, et al. Roformer: enhanced transformer with rotary position embedding[EB/OL]. (2021-04-20) [2021-12-01]. https://arxiv. org/pdf/2104. 09864.

[22] SU J L, ZHU M, MURTADHA A, et al. ZLPR: a novel loss for multi-label classification[EB/OL]. (2022-08-05) [2022-09-12]. https://arxiv.org/pdf/2208.02955.

[23] PENG N, DREDZE M. Improving named entity recognition for chinese social media with word segmentation representation learning[EB/OL]. (2016-03-02) [2021-12-05].https://arxiv.org/pdf/1603. 00786.

[24] ZHANG N, CHEN M, BI Z, et al. Cblue: a chinese biomedical language understanding evaluation benchmark [EB/OL]. (2021-06-15) [2021-12-24]. https://arxiv.org/pdf/2106.08087.

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

[26] 毛明毅, 吴晨, 钟义信, 等. 加入自注意力机制的 BERT 命名实体识别模型[J]. 智能系统学报, 2020, 15(4): 772-779.
MAO Mingyi, WU Chen, ZHONG Yixin, et al. BERT named entity recognition model with self-attention mechanism[J]. Journal of Intelligent Systems, 2020, 15(4): 772-779.

[27] 李健, 熊琦, 胡雅婷, 等. 基于 Transformer 和隐马尔科夫模型的中文命名实体识别方法[J]. 吉林大学学报(工学版),2023, 53(5):1427-1434.
LI Jian, XIONG Qi, HU Yating, et al. Chinese named entity recognition method based on Transformer and hidden Markov model[J]. Journal of Jilin University (Engineering Edition), 2023, 53(5):1427-1434.

[28] 钟诗胜, 陈曦, 赵明航, 等. 引入词集级注意力机制的中文命名实体识别方法[J]. 吉林大学学报(工学版), 2022, 52(5): 1098-1105.
ZHONG Shisheng, CHEN Xi, ZHAO Minghang, et al. Chinese named entity recognition method based on word set level attention mechanism[J]. Journal of Jilin University (Engineering Edition), 2022, 52(5): 1098-1105.

(编辑:郭少华)