doi:10.3969/j.issn.1672-1292.2024.04.008

基于边界信息和词汇信息增强的中文命名实体识别

孙争艳[1],陈　磊[1],魏苏波[2],陈宝国[1]

(1.淮南师范学院计算机学院,安徽 淮南 232038)
(2.上海大学计算机工程与科学学院,上海 200444)

[**摘要**]　中文命名实体识别(named entity recognition,NER)是一种提取实体对的自然语言处理(natural language processing,NLP)技术,广泛应用于知识图构建和信息提取任务中. 传统的中文 NER 方法主要强调字符信息的分析,而忽略了位置和单词特征等重要方面,阻碍了实体边界的准确识别. 引入了一种增强的中文命名实体识别模型,该模型高度重视边界和单词信息,以实现实体边界的精确校准. 首先,构建多层次文本特征作为模型的输入. 然后,提出了融合位置信息和类别描述信息的策略,以增强语义表示能力. 最后,使用条件随机场模型将增强的特征向量映射到序列标签输出,以准确提取所有实体和类别标签. 模型在现有数据集 OntoNotes、Resume 和 Weibo 上,$F1$ 得分分别提高了 0.82%、0.78%和 1.51%,验证了模型的有效性.

[**关键词**]　命名实体识别,位置信息,类别描述信息,多层次文本特征
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1672-1292(2024)04-0079-08

Named Entity Recognition Based on Boundary Information and Word Information Enhancement

Sun Zhengyan[1],Chen Lei[1],Wei Subo[2],Chen Baoguo[1]

(1.College of Computer,Huainan Normal University,Huainan 232038,China)
(2.School of Computer Engineering and Science,Shanghai University,Shanghai 200444,China)

Abstract:Chinese named entity recognition(NER) is a natural language processing(NLP) technology that extracts entity pairs,which is widely used in knowledge graph construction and information extraction tasks. The traditional Chinese NER method mainly emphasizes character-level analysis,but ignores important aspects such as location and word features,which hinders the accurate identification of entity boundaries. This paper introduces an enhanced Chinese NER model that places a heightened emphasis on both boundary and word information to enable the precise calibration of entity boundaries. Firstly,multi-level text features are constructed as the input of the model. Then,the strategy of integrating location information and category description information is proposed to enhance the semantic representation ability. Finally,the conditional random field(CRF) model is used to map the enhanced feature vector to the serialized label output to accurately extract all entity and category labels. The efficacy of the proposed model is underscored by empirical evidence,revealing advancements in the $F1$ score by increments of 0.82%,0.78%,and 1.51% on the existing datasets OntoNotes,Resum and Weibo,respectively.

Key words:named entity recognition,location information,category description information,multi-level text features

伴随着深度学习模型的不断涌现,命名实体识别(named entity recognition,NER)已在信息检索、知识图谱构建等多个领域中扮演着至关重要的角色. NER 的核心任务在于从文本中精准地识别出个人名字、地名[1]等关键信息,并对这些实体类别进行系统的分类. 相较于英文文本,中文文本在结构上展现出了更加复杂和多样的特点. 首先,英文的单词间通常以空格作为辨识界限,相比之下,中文文本的词汇边界并

收稿日期:2024-05-12.
基金项目:安徽省科研计划编制项目重点项目(2024AH051731)、国家重点实验室开放基金项目(COGOS-2023HE02)、淮南市指导性科技计划项目(4302)、淮南师范学院校级专项重点(基础教育)项目(2023XJZD025).
通讯作者:陈磊,硕士,教授,研究方向:数据挖掘、实体识别. Email:leichen@ hnnu.edu.cn

不明显,词语的界定往往需要依赖上下文和分词技术进行识别. 以图 1 为例,任务是从文本中精确辨认出实体“张爱玲”. 若仅以单一的字符信息作为参考,难以将之视为一个完整的字符序列. 若考虑分词技术,则需要能准确划分出“张爱玲”这一实体.

其次,中文词汇的词性与含义具有丰富多变性. 如图 1 所示,“张爱玲”由姓“张”及名“爱玲”组成,表示特定人名. 若将“张”字用于“张开”一词时,也指代动作. 可见,中文字在不同组合及上下文中,其含义与词性各异. 这一特点要求在理解和处理中文文本时须准确掌握字词的具体语境. 此外,中文特有的复杂句子结构侧重于语义联系而非语法结构. 如图 1 所举例子中,“张”作为常见姓氏,“爱玲”作为常见名字,两者之间存在共现性与语义联系. 若模型能学习并掌握此种关联性,将有助于将“张爱玲”正确识别为一个整体的人名实体,而避免将其错误拆分为多个独立实体.

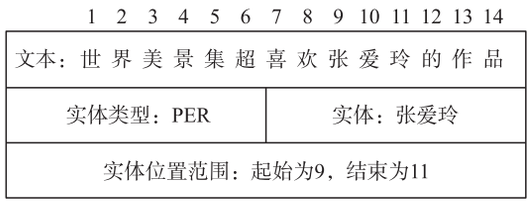

图 1　命名实体识别实例

Fig. 1　Named entity recognition instance

现有的传统模型已注意到这些问题,然而研究者们往往聚焦于字符层面的分析,却忽视了位置信息和词汇信息的重要性. 这种忽略不可避免地影响了实体边界识别的精确度,继而对整体的实体识别准确性产生了不良效果. 鉴于上述考虑,本文提出了一种融合边界信息与词汇信息增强的中文 NER 模型. 该模型不仅着力于字符和词汇信息的学习,还结合类别描述信息以强化类别学习,并在字符信息处理中融入位置信息,且使用实体跨度来界定实体的边界,以此提升实体边界识别的精度. 本文中的模型框架如图 2 所示.

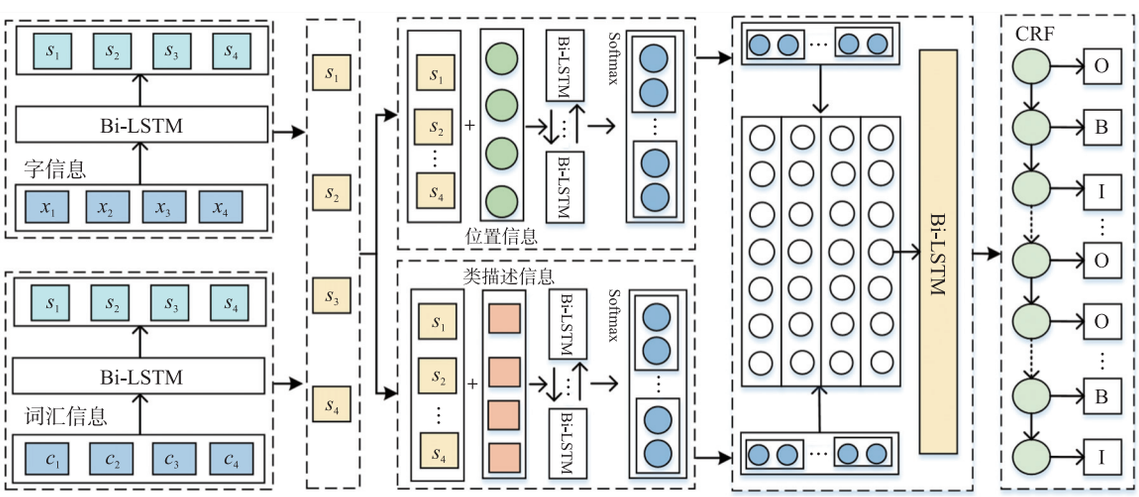

图 2　模型框架图

Fig. 2　Diagram of model frame

首先,独立学习字信息与词汇信息,以增强语义表征能力. 对每个时间步长的字符嵌入和词汇嵌入进行加权求和,将加权求和后的向量作为融合后的输入,传入后续模型层. 其次,将类描述信息融入文本中,以增强类别标签信息的学习能力,进一步提升对标签语义信息的掌握. 针对所得向量,整合了位置信息,包括候选实体的起始和终止位置,并以跨度形式表示. 基于此,通过多层感知器层计算每个单词作为实体首尾单词的概率.

1　相关工作

1.1　基于规则和统计的方法

基于规则的 NER 方法主要通过正则表达式和词典进行实体匹配. 正则表达式凭借特定字符组合来验证文本,而词典则基于实体集合. 此外,设计规则主要用于精确或模糊匹配,在中文环境下,此方法需设计人名、地名、机构等实体[1]的规则,且规则比英文更复杂. 在构造规则时,关键是规则的数量和质量. Collins 等[2]建立了一个预定义的种子规则集 DecisionList,然后通过对语料库应用无监督训练迭代,获得更多规

则,最终生成大量规则来识别实体. Cucerzan 等[3]提出了一种自动生成规则的方法,提高了工作效率. 基于规则的方法在特定对象上可以获得良好的性能,然而,手动构造规则库既需要大量专业知识,也需要耗费大量人力. 在此基础上,基于统计的方法开始得到普遍应用.

基于统计的方法将 NER 任务转换为序列标记任务. 其中,基于机器学习的统计方法被广泛应用,主要包括隐马尔可夫模型[4]、最大熵[5]、条件随机场[6]和支持向量机[7]等. 这些模型能够通过非线性激活函数学习更复杂的特征,并自动发现潜在的特征. 然而,这些方法仍然需要构造特征工程,于是开始尝试新的思路,并逐渐转向深度学习方法以寻求更好的解决方案.

1.2 基于深度学习的方法

基于深度学习的 NER 方法通常使用编码器-解码器结构,主要包括预处理、特征提取和解码技术. 预训练技术通过大量未标注文本训练深度网络,从早期的 Word2Vec[8]到现有的动态模型如 BERT[9]和 ALBERT[10]均采用了这种技术. 特征提取环节目标为获取上下文向量表示和语义信息,在解码阶段通常采用条件随机场完成,以输出高全局概率的实体标记序列.

针对中文 NER 任务,提出了许多方法以应对特定问题. 例如,孙振等[11]针对医疗文本中的类型分布不均衡问题,提出了多特征融合模型,以增强语义联系. 雷松泽等[12]通过结合外部知识实现了多特征融合. 韩晓凯等[13]从语义信息出发,提出了注意力增强模型筛选重要信息. 崔少国等[14]从实体边界入手,利用笔画特征信息来丰富语义信息. 宋旭晖等[15]通过协同图网络将分词信息融入模型来充分利用边界信息. Chen 等[16]提出了一个统一的框架来学习边界信息,以更好地进行中文命名实体识别. Gui 等[17]提出了一种基于卷积神经网络的中文 NER 方法,采用重新思考机制来整合词典,解决了 GPU 并行化和词典冲突的问题. 近年来,应用深度学习进行 NER 的趋势更加倾向于集成学习[18]、迁移学习[19]和对抗性学习[20]等方法,同时涌现出越来越多的新模型.

2 基于位置信息增强的实体识别方法

2.1 实体识别任务基本定义

给定输入序列 $x=\{x_1,\cdots\cdots,x_i\}$,其中 $i=2,\cdots\cdots,L$. 通过序列标注方法得到标签输出序列 $y=\{y_1,\cdots,y_i\}$,其中 y_i 表示 x_i 的预测标签,y 中包括所有预定义实体类型集与 O,即 $y\in\{Y\cup O\}$,其中 Y 是预定义的实体类集合,O 表示非实体类型.

由于中文 NER 任务多数序列标注质量都不高,且存在标注错误的数据,因此本文从源数据集中标记的数据构建训练集 $E_{\text{train}}=\{(T_{\text{train}},A_{\text{train}},L_{\text{train}})\}$,其中 T_{train}表示仅包含数据文本的集合,L_{train}表示信息增强的数据集,包括文本、实体类型以及类描述信息,A_{train}表示包含所有数据的集合,即 $A_{\text{train}}=\{T_{\text{train}}\cup L_{\text{train}}\}$. 在测试阶段,以类似的方式用来构建数据集.

2.2 字信息和词汇信息的编码

给定一个包含 n 个“token”的字 $x=\{x_1,\cdots,x_n\}$,“token”表示为

$$[s_1,\cdots,s_i]=\text{BERT}(x_1,\cdots,x_i). \tag{1}$$

BERT 是包含语境信息的预训练模型,旨在更好地表示和预测文本的跨度. 首先,使用不同的随机过程来掩盖“token”,使用其边界标记“#”分隔实体,同样地,对于词汇信息进行“token”表示.

$$[s_1,\cdots,s_j]=\text{BERT}(c_1,\cdots,c_j). \tag{2}$$

其中,$\{s_1,\cdots,s_n\}$表示序列中每个“token”的输出. 给定“token”的掩码范围(s_t,s_e),其中 s_t 和 s_e 分别表示词汇开始和结束位置,使用词汇的外边界 s_{t-1}和 s_{e+1},以及相对位置信息 p_{i-t+1}求得隐藏向量,BERT 的 2 层前馈网络表示为

$$\boldsymbol{h}_0=f(s_{t-1},s_{e+1},p_{i-t+1}), \tag{3}$$

$$\boldsymbol{h}_1=\text{LayerNorm}(\text{GeLU}(\boldsymbol{W}_1\boldsymbol{h}_0)), \tag{4}$$

$$\boldsymbol{h}_2=\text{LayerNorm}(\text{GeLU}(\boldsymbol{W}_2\boldsymbol{h}_1)). \tag{5}$$

其中,位置嵌入 $p_1,p_2,\cdots,p_{i-t+1}$表示“tokens$_i$”的相对位置,使用 GeLU 作为激活函数,使用向量表示 $\boldsymbol{h}_i$ 来预测词汇向量. $\boldsymbol{W}_1$ 表示第 1 层神经网络权重矩阵,$\boldsymbol{W}_2$ 表示第 2 层神经网络的权重矩阵,权重矩阵主要由损失函数的梯度逐步调整和更新.

2.3　嵌入位置信息的边界增强模型

边界检测旨在识别一个词是实体的第一个词还是最后一个词. 本文使用两个“token”分类器预测实体的开始和结束位置. 具体地,本文将上下文表示 $\boldsymbol{h}_i$ 输送到多层感知器分类器中,并应用 softmax( )激活函数来获得单词 w_i 作为实体的第一个单词的概率 P_t^i.

$$P_t^i=\text{softmax}(\text{forward}(\boldsymbol{h}_i)). \tag{6}$$

类似地,应用多层感知器分类器来获得单词 w_i 作为实体最后一个单词的概率 P_e^i.

$$P_e^i=\text{softmax}(\text{forward}(\boldsymbol{h}_i)). \tag{7}$$

式中,forward( )表示多层感知器的前向传播计算函数,softmax( )表示激活函数,将输入向量转换为概率分布,表示输入属于每个类别的概率. 在训练过程中,由于每个句子可能包含多个实体,甚至存在实体嵌套问题,实体跨度的边界检测尤为重要,因此本文将所有实体的跨度边界标记作为重要位置信息嵌入模型中,本文将实体边界训练目标函数定义为检测过程中两个后续交叉损失的总和.

$$L_b^t=-\sum_{i=1}^{N}\left[y_t^i\ln P_t^i+(1-y_t^i)\ln(1-P_t^i)\right], \tag{8}$$

$$L_b^e=-\sum_{i=1}^{N}\left[y_e^i\ln P_e^i+(1-y_e^i)\ln(1-P_e^i)\right], \tag{9}$$

$$L_b=L_b^t+L_b^e. \tag{10}$$

其中,y_t^i 和 y_e^i 分别表示单词 i 为实体的第一个和最后一个单词的标签信息,P_t^i 和 P_e^i 表示单词 i 为实体的开始“token”和结束“token”的概率. 最后,将最大跨距表示 h_p^i 与外边界表示(s_{t-1},s_{e+1})连接起来表示候选实体边界$(\boldsymbol{h}_{t_{i-1}},\boldsymbol{h}_{e_{i+1}})$,用于预测实体的边界表示.

3　基于类描述信息增强的实体识别方法

3.1　BIO 标注

数据预处理主要任务包含两个方面,一是对文本的内容进行初步筛选,对句子成分缺省删除或者是对句子格式进行统一,二是对筛选后的句子进行标注. 首先,对一些无意义的符号移除或替换,使用 Jieba 工具进行分词和词性标注. 词信息标注通常包括位置标注(如 BIO 标注法:B 表示词的开头,I 表示词的中间,O 标记不属于任何实体的部分)和实体类别标注(如 PER 表示人名,ORG 表示组织名,LOC 表示地名等). 有时分词或标注工具会产生错误. 例如,把“北京大学”错误分成“北京/大学”. 此时需要人工校正或通过规则来自动修正.

3.2　类标签信息和描述信息的嵌入

本文提出了一种基于类描述信息增强的实体识别方法,旨在提高 NRE 模型的精度和鲁棒性. 具体地,首先,通过对字符和词汇信息进行编码,形成基本信息的表示. 然后,将类标签信息与类描述信息作为增强信息添加到基本信息中. 通过将基本信息和增强信息联合编码,模型可以更好地捕捉实体的上下文和类别信息.

在该方法中,基本信息是经过编码的输入句子,而增强信息则包括类描述和标签信息,通过加入的增强信息构建增强信息数据集,该数据集帮助模型更精确地识别和区分不同类型的实体. 所构建数据集用于模型的训练,其中训练阶段利用基本信息和增强信息以及由 softmax( )激活函数构成的分类器,确保所识别的实体及其类别在原句中得到清晰展现. 这样的增强信息模型应实现较高的识别精度和对实体类别辨识的鲁棒性.

如果将 m^t 表示为步骤 t 的训练模型. m^t 从 m^{t-1}学习保存旧类别的知识,主要是旧类的描述信息,从旧类库 D^t 中学习旧类信息或者生成新的类信息,同时保存 t 前的实体类别的信息$\{D^k\}_{k=1}^{t-1}$,将类描述信息和标签信息嵌入合成新的数据集 $D_r^t=\{E_i^{t,r},Y_i^{t,r}\}_{i=1}^{|D_t^t|}$,其中 $E_i^{t,r}$ 和 $Y_i^{t,r}$ 分别是类描述信息嵌入和参考标签序列嵌入.

在增强信息中有描述信息,用符号[MS]开头,后面是所有类及其描述信息的列表,$Z_{ms}=\{([MS],L_1,L_2,\cdots,L_i)\mid L_i\in S\}$,$S$ 是指描述实体类型的句子的“token”集合,$L_i=(l_1^1,l_1^2,\cdots,l_1^i)$,$l_i^j$ 是第 i 种类型的第 j 个概念. 然后再加上实体和对应的类型,以 $e_i=ty_i$ 的格式生成文本,e_i 是实体的向量表示,ty_i 是实体对应类

型的向量表示,用于后续的特征提取. 通过从 m^{t-1} 中提取信息来训练 m^t,使用 D^t 的真实数据和 D_r^t 的合成数据训练. 训练目标将所有特征进行融合,使得损失函数最小,从而优化模型参数,进而提高模型精度. 具体的算法如表 1 所示.

表 1　命名实体识别算法描述

Table 1　Named entity recognition algorithm description

算法　命名实体识别算法
输入:文本 $x=\{x_1,\cdots,x_n\}$,词汇信息 $c=\{c_1,\cdots,c_n\}$,类标签信息及类描述信息 Z_{ms}; 输出:$Y=(y_1,y_2,\cdots,y_n)$. (1) NamedEntitInput(x,c,$\boldsymbol{Z}_{ms}$);/ * 数据预处理获取初始输入数据集,包括字符信息、词汇信息以及增强信息 * / (2) for each text x (3) Si←BERT(x);/ * 将文本中的每个字符使用 BERT 进行编码得到 Si 表示 * / (4) end for (5) for each text x (6) Si←BERT(x,c);/ * 将文本中的每个词汇使用 BERT 进行编码,将加权求和后的向量作为融合后的输入 Si * / (7) end for (8) $\boldsymbol{hi}$←Bi-LSTM(w,Si);/ * 将边界跨度位置信息编码融入 Si 中,生成位置增强后的向量表示 $\boldsymbol{hi}$ * / (9) $\boldsymbol{hi}$←Bi-LSTM(e,Si);/ * 将类及类描述信息信息编码融入生成 $\boldsymbol{hi}$ * / (10) Feature fusion (11) Pi←Softmax(forward($\boldsymbol{hi}$));/ * 获取候选实体概率矩阵 * / (12) Fi=getFeature($\boldsymbol{hi}$,($\boldsymbol{h}_{t_{i-1}}$,$\boldsymbol{h}_{e_{i+1}}$)) (13) $Y(y_1,y_2,\cdots\cdots,y_n)=CRF(Fi)$ (14) end

4　实验

4.1　数据集及评价标准

本文选用 3 个不同数据集上实验的数据:OntoNotes[21]、Resume[22]和 Weibo[23].

OntoNotes:包含 170 万英语单词、100 万中文词和 30 万阿拉伯语词,覆盖新闻、通话、网络博客、广播和脱口秀等多个领域,涉及 18 种实体类型.

Resume:该数据集由上市公司高管简历构成,共收集了 1 027 份简历摘要,并由 YEDDA 手动标注了国家、位置、人名等 8 种实体.

Weibo:该数据集源自新浪微博,涵盖 2013 年 11 月至 2014 年 12 月的 1 890 条社交媒体消息样本. 数据集聚焦 4 个语义实体类别:个人、组织、地点和地区,并采用多种标注手段,同时对两种类别的名称以及名词性代词提及进行了注解.

以上数据集的相关信息如表 2 所示,由于数据集数据量大,本文使用 80%、10%、10%的训练、测试、验证集进行分割.

表 2　实验数据集来源

Table 2　Source of experimental dataset

数据集	种类	网址	训练集/个	测试集/个	验证集/个
OntoNotes	18	https://catalog.ldc.upenn.edu/LDC2013T19	10 560	1 320	1 320
Resume	8	https://github.com/jiesutd/LatticeLSTM	823	102	102
Weibo	4	https://github.com/hltcoe/golden-horse/tree/master/data	1 512	189	189

目前主流的方法是根据类型、预测的标签是否正确对 NER 性能进行评分. 本文认为如果输出实体的类型和边界正确,则输出实体是正确的,实体的边界和实体类型是准确的. 对于每个类别,主要是通过混淆矩阵得到精确率 P、召回率 R 和 $F1$ 值.

$$P=\frac{TP}{TP+FP}\times 100\%, \tag{11}$$

$$R=\frac{TP}{TP+FN}\times 100\%, \tag{12}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\%. \tag{13}$$

4.2　对比实验

为了验证本文提出的模型对实体识别任务的有效性,在上文的数据集和评价标准下进行了对比实验,将基于边界信息和词汇信息增强的中文 NER 模型与其他模型进行对比,以验证模型的性能.

Star+GAT+MultiTask[16]:提出了一种边界信息增强策略. 该策略一方面通过图注意力网络层加强短语依存关系表示. 另一方面,在统一框架内结合实体头尾预测作为辅助任务,共同优化边界信息的学习与实体的识别.

LR-CNN[17]:以卷积神经网络为基础,引入重新思考机制对词典进行整合,能够同时对所有字符及句中可能匹配的词汇进行并行建模.

Lattice LSTM[21]:该模型接受字符序列及所有潜在匹配的词典词汇输入,有效结合字与词信息,克服分词错误.

TL-NER[22]:融合迁移学习的深度学习模型,从大规模未标注文本中提取信息来避免分词误差,实现高效的知识迁移.

SDI-NER[23]:提出了一种新的 NER 句法依存图信息学习模型,并从中提取各种特定任务的隐藏信息多个 CWS 和词性标记任务,以进一步改进 NER 模型.

LSTM-CRF[24]:NER 的经典基线模型.

CNA-NER[25]:该模型结合局部注意力卷积神经网络和全局自我关注 Bi-GRU,以提取词级特征和字符级的上下文信息.

BGRU-CRF[26]:该模型通过 BGRU-CRF 融合上下文信息,减少实体认识的歧义,并结合注意力机制筛选关键字词,捕捉长距离依赖,以准确执行 NER 任务.

W^2NER[27]:W^2NER 模型通过下一个相邻词和尾头词方法分析实体间的邻接关系,形成了一个神经框架,将统一的 NER 描述为由词对组成的 2D 网格.

根据上述模型与本文模型进行对比,通过 P 值、R 值、$F1$ 值比较具体内容如表 3、表 4 和表 5 所示.

表 3　在数据集 OntoNotes 的对比实验

Table 3　Comparative experiment on dataset OntoNotes

模型	P	R	$F1$	模型	P	R	$F1$
Star+GAT+MultiTask	79.95	79.95	79.95	LSTM-CRF	68.14	66.47	67.29
LR-CNN	74.45	74.45	74.45	CNA-NER	75.05	72.29	73.64
Lattice LSTM	75.64	76.91	76.27	BGRU-CRF	77.62	75.34	76.46
TL-NER	76.56	70.27	73.28	W^2NER	83.08	83.08	83.08
SDI-NER	76.96	76.94	76.95	本文模型	84.12	83.28	83.70

表 4　在数据集 Resume 的对比实验

Table 4　Comparative experiment on dataset Resume

模型	P	R	$F1$	模型	P	R	$F1$
Star+GAT+MultiTask	96.08	93.45	94.75	LSTM-CRF	88.16	87.43	87.79
LR-CNN	95.11	95.11	95.11	CNA-NER	93.25	92.18	92.71
Lattice LSTM	92.78	92.35	92.56	BGRU-CRF	95.76	95.04	95.39
TL-NER	95.75	93.89	94.81	W^2NER	96.65	96.65	96.65
SDI-NER	96.04	95.86	95.95	本文模型	96.82	96.98	96.90

表 5　在数据集 Weibo 的对比实验

Table 5　Comparative experiment on dataset Weibo

模型	P	R	$F1$	模型	P	R	$F1$
Star+GAT+MultiTask	69.50	69.50	69.50	LSTM-CRF	67.79	68.14	67.96
LR-CNN	59.92	59.92	59.92	CNA-NER	68.05	69.97	69.00
Lattice LSTM	67.25	66.84	67.04	BGRU-CRF	71.02	69.36	70.18
TL-NER	69.46	67.83	68.64	W^2NER	72.32	72.32	72.32
SDI-NER	73.61	54.08	62.94	本文模型	74.24	73.42	73.83

从表 3 至表 5 的分析显示,本文模型在性能上胜过 Star+GAT+MultiTask 模型,原因是后者虽加强边界识别,却忽略了标签信息的作用. 相对于 LR-CNN,本模型的优势在于采用了跨度表示增强边界方法,通过

双重分类器提升实体边界判定的准确性. 此外,本模型更有效地利用了标签信息,而 Lattice LSTM 在这方面有所不足. TL-NER 模型虽采用转移学习来应对知识迁移,但在处理跨领域数据集如 OntoNotes 时不尽人意,未充分考虑特定领域信息. 本文模型较 SDI-NER 模型显示出优越性,不仅考虑局部关键信息和位置数据,还融入了全局信息增强了对实体上下文语义的表征能力. 经典模型 LSTM-CRF 的性能相对较弱,这主要是因为该模型未充分考虑词汇与位置信息对实体识别的重要性,并且在区分词汇与字符信息方面存在不足,进而削弱了识别准度. 相较于 CNA-NER 模型,本文模型有所进步,不仅捕捉单词和字符特征,还结合了标签与位置信息优化实体边界识别. BGRU-CRF 能处理长距离依赖问题,本文模型在此基础上进一步提取隐特征向量,改善了实体边界与上下文信息的整合. W^2NER 模型虽增强了 NER 中的语义表示,却忽视了类标签和描述信息的价值.

4.3 消融实验

本研究中各模块具有特定分工,为检验其有效性,在数据集 Resume 和 Weibo 上进行消融实验. 实验旨在验证各主要模块的重要性. 表 6 和图 3 中的符号定义为:Overall 表示完整模型,-CH 代表没有词汇信息嵌入,-WZ 代表无位置信息增强,-BQ 代表无标签信息嵌入,-LSM 代表无类别描述增强. 具体于数据集 Resume 的实验结果见表 6.

表 6 在数据集 Resume 的消融实验

Table 6 Ablation experiment on dataset Resume

模型	*P*	*R*	*F*1
Overall	96.82	96.98	96.90
-CH	95.56	94.37	94.96
-WZ	87.48	85.91	86.69
-BQ	92.53	90.94	91.73
-LMS	91.64	89.85	90.74

现将各个模块在数据集 Weibo 的 *P* 值、*R* 值和 *F*1 值绘制成柱形图,具体结果如图 3 所示. 横坐标代表缺失的模块,纵坐标代表相应模型的数据集数值.

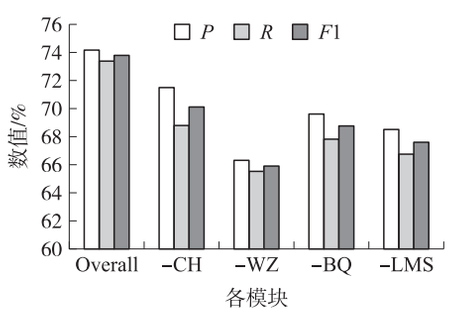

图 3 在数据集 Weibo 的消融实验

Fig. 3 Ablation experiment on dataset Resume

根据表 6 和图 3 的数据,位置信息增强模块的缺失比其他模块更影响实体识别,表示边界信息对实体识别至关重要. 相比有类别描述信息,无类别描述增强模块在数据集 Resume 和 Weibo 上分别减少了 6.16%和 6.15%. 消融实验进一步证实了词汇信息嵌入模块的有效性,特别是在数据集 Resume 上,*F*1 值提高了 1.94%. 同时,标签信息嵌入模块也得到验证,显著提高了精确率和召回率. 因此,研究采用的各模块均有效提升了实体识别性能,为中文 NER 贡献了切实价值.

5 结论

中文文本由于其独特性及灵活性,若上下文语义关系表征不明确会降低准确率. 单独学习字符向量会忽视词汇信息特征,而仅依赖词汇信息可能因分词错误引发更大的分类误差. 因此,本文将字符与词汇信息相结合,以挖掘更多潜在特征.

为解决实体边界定义模糊的问题,本研究在实体识别模型中引入了一种基于位置信息的边界增强方法. 该方法计算文本中每个“token”的边界概率,以此明确每个实体的跨度. 其次,提出了一个融合类标签信息和类描述信息的增强数据集,解决了类标签信息缺乏的问题. 最后,利用条件随机场实现序列化的输出. 本文通过对比实验和消融实验两种方式验证了模型的有效性. 考虑到现有中文实体识别标注数据集较少,未来研究可采用无监督方法实现标注工作的批量化,有望在信息检索和问答系统等领域获得更广泛应用.

[参考文献](References)

[1] 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.

[2] COLLINS M,SINGER Y. Unsupervised models for named entity classification[C]//Proceedings of the 1999 Joint SIGDAT

Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. College Park,MD,USA,1999.
[3] CUCERZAN S, YAROWSKY D. Language independent named entity recognition combining morphological and contextual evidence[C]//Empirical Methods in Natural Language Processing. 1999.
[4] LI Y,SONG L,ZHANG C. Sparse conditional hidden Markov model for weakly supervised named entity recognition[C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York,NY,US:Association for Computing Machinery,2022:978-988.
[5] LIU P,GUO Y M,WANG F L,et al. Chinese named entity recognition:The state of the art[J]. Neurocomputing,2022,473: 37-53.
[6] AN Y,XIA X Y,CHEN X L,et al. Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF[J]. Artificial Intelligence in Medicine,2022,127:102282.
[7] GOVINDARAJAN S, MUSTAFA M A, KIYOSOV S, et al. An optimization based feature extraction and machine learning techniques for named entity identification[J]. Optik,2023,272:170348.
[8] LIU Y X,WANG L,SHI T F,et al. Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM[J]. Information Systems,2022,103:101865.
[9] ELANGOVAN A,LI Y,PIRES D E V,et al. Large-scale protein-protein post-translational modification extraction with distant supervision and confidence calibrated BioBERT[J]. BMC Bioinformatics,2022,23(4):1-23.
[10] CHEN M J,LUO X,SHEN H L,et al. A novel named entity recognition scheme for steel e-commerce platforms using a lite BERT[J]. Computer Modeling in Engineering & Sciences,2021,129(1):47-63.
[11] 孙振,李新福. 多特征融合的中文电子病历命名实体识别[J]. 计算机工程与应用,2023,59(23):1-10.
[12] 雷松泽,刘博,王瑜菲,等. 结合多特征嵌入和多网络融合的中文医疗命名实体识别[J]. 电子与信息学报,2023,45(8):1-8.
[13] 韩晓凯,岳颀,褚晶,等. 基于注意力增强的点阵 Transformer 的中文命名实体识别方法[J]. 厦门大学学报(自然科学版),2022,61(6):1062-1071.
[14] 崔少国,陈俊桦,李晓虹. 融合语义及边界信息的中文电子病历命名实体识别[J]. 电子科技大学学报,2022,51(4): 565-571.
[15] 宋旭晖,于洪涛,李邵梅. 基于图注意力网络字词融合的中文命名实体识别[J]. 计算机工程,2022,48(10):298-305.
[16] CHEN C,KONG F. Enhancing entity boundary detection for better 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. Online,2021:20-25.
[17] GUI T,MA R T,ZHANG Q,et al. CNN-Based Chinese NER with lexicon rethinking[C]//Twenty-eighth International Joint Conference on Artificial Intelligence. Macao,China,2019:4982-4988.
[18] 梁兵涛,倪云峰. 基于集成学习的中文命名实体识别方法[J]. 南京师大学报(自然科学版),2022,45(3):123-131.
[19] 吴炳潮,邓成龙,关贝,等. 动态迁移实体块信息的跨领域中文实体识别模型[J]. 软件学报,2022,33(10):3776-3792.
[20] 孔令巍,朱艳辉,张旭,等. 基于对抗训练的中文电子病历命名实体识别[J]. 湖南工业大学学报,2022,36(3):36-43.
[21] ZHANG Y,YANG J. Chinese NER using lattice LSTM[J]. arXiv Preprint arXiv:1805.02023,2018.
[22] PENG D L, WANG Y R, LIU C, et al. TL-NER: A transfer learning model for Chinese named entity recognition[J]. Information Systems Frontiers,2020,22(6):1291-1304.
[23] ZHU P,CHENG D W,YANG F Z,et al. Improving Chinese named entity recognition by large-scale syntactic dependency graph[J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing,2022,30:979-991.
[24] CHEN T Y,HU Y M. Entity relation extraction from electronic medical records based on improved annotation rules and BiLSTM-CRF[J]. Annals of Translational Medicine,2021,9(18):1415.
[25] ZHU Y Y,WANG G X. CAN-NER:Convolutional attention network for Chinese named entity recognition[J]. arXiv Preprint arXiv:1904.02141,2020.
[26] 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.
[27] LI J Y,FEI H,LIU J,et al. Unified named entity recognition as word-word relation classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence. 2022,36(10):10965-10973.

[责任编辑:陈　庆]