·开发研究与工程应用·　　文章编号：1000-3428(2022)08-0313-08　　文献标志码：A　　中图分类号：TP391

用于嵌套命名实体识别的边界强化分类模型

连艺谋，张英俊，谢斌红

（太原科技大学 计算机科学与技术学院，太原，030024）

摘　要：实体嵌套是自然语言中一种常见现象，提高嵌套命名实体识别的准确性对自然语言处理各项任务具有重要作用。针对现有嵌套命名实体识别方法在识别实体边界时不够准确、未能有效利用实体边界信息等问题，提出一种嵌套命名实体识别的边界强化分类模型。采用卷积神经网络提取邻接词的特征，通过加入多头注意力的序列标注模型获取实体中的边界特征，提高实体边界检测的准确性。在此基础上，计算实体中各词语对实体类型的贡献度，将实体关键字与实体边界词相结合来表示实体，使实体表示中包含实体关键信息和边界信息，最后进行实体类型检测。实验结果表明，通过加入多头注意力机制能够有效提升对嵌套命名实体的检测和识别性能，该模型在GENIA和GermEval 2014数据集上准确率有较好表现，并且召回率和F1值较对比模型达到最优。

关键词：嵌套命名实体识别；实体表示；注意力机制；边界；神经网络

开放科学（资源服务）标志码（OSID）：

中文引用格式：连艺谋，张英俊，谢斌红.用于嵌套命名实体识别的边界强化分类模型[J].计算机工程，2022，48(8)：313-320.

英文引用格式：LIAN Y M，ZHANG Y J，XIE B H.Boundary enhanced classification model for nested named entity recognition[J].Computer Engineering，2022，48(8)：313-320.

Boundary Enhanced Classification Model for Nested Named Entity Recognition

LIAN Yimou，ZHANG Yingjun，XIE Binhong

（School of Computer Science and Technology，Taiyuan University of Science and Technology，Taiyuan 030024，China）

【Abstract】Entity nesting is a common phenomenon in natural language.Improving the accuracy of nested Named Entity Recognition（NER）plays an important role in various Natural Language Processing（NLP）tasks.Addressing the inaccuracy of existing nested NER methods for identifying entity boundaries and their ineffective use of entity boundary information，a boundary enhanced classification model for nested NER is proposed.A Convolution Neural Network（CNN）is used to extract features of adjacent words，and sequence annotation model with multi-heads attention is added to obtain the boundary features of entities and improve the accuracy of entity boundary detection.On this basis，the model first calculates the contribution of each word in the entity to each entity type，combines the entity keyword with each entity boundary word to represent the entity，makes the entity representation contain the entity key and boundary information，and finally determines the entity type.Experiments show that adding the multi-head attention mechanism effectively improves the performances of nested NER and recognition.The model demonstrated a good accuracy performance on the GENIA and GermEval 2014 datasets.In addition，of all models compared in the experiments，the proposed model achieved the best recall rate and F1-score.

【Key words】nested Named Entity Recognition（NER）；entity representation；attention mechanism；boundary；neural network

DOI：10.19678/j.issn.1000-3428.0062181

0　概述

命名实体识别（Named Entity Recognition，NER）[1]是自然语言处理（Natural Language Processing，NLP）中的一项基本任务，目的是识别出自然语言文本中具有特定意义的语言块，如人名、地名、组织名等。命名实体识别的结果会作为前期基础数据输入到关系抽取、事件抽取、知识图谱等下游任务中，其识别的准确性一定程度上决定了NLP应用的最终效果。在实体中可能会存在嵌套实体，例如“IL-2 Promoter”是一种“DNA”实体，该实体中的“IL -2”又是一种“protein”实体。根据实体中是否存在嵌套实体的情况，NER可分为flat NER

基金项目：山西省重点研发计划（重点）高新领域项目（201703D111027）；山西省重点研发计划（201803D121048，201803D121055）。

作者简介：连艺谋（1996—），男，硕士研究生，主研方向为深度学习、自然语言处理；张英俊，教授；谢斌红，副教授。

收稿日期：2021-07-26　　**修回日期：**2021-09-21　　**E-mail：**1484883694@qq.com

和nested NER。实体嵌套是自然语言处理中的一种常见问题，研究该问题的解决方案，对落实命名实体识别的应用具有重要意义。

目前对于NER的研究大多是针对flat NER的研究，然而在医学、新闻等领域的数据集中存在大量嵌套实体，如生物医学领域常用的GENIA语料库中存在约10%的嵌套实体。在处理flat NER时，现有方法通常是将其当作序列标注的问题去解决[2-3]，如BiLSTM-CRF模型[4]。但是这种方法要求每个词在标注时对应一个标签，而在nested NER中，每个词会对应多个标签，将flat NER方法直接应用在nested NER上效果并不理想。因此，解决nested NER存在的问题，提高命名实体识别的效果，是当前的研究热点。

对于nested NER任务，一种解决方法是借助超图的方法，该方法以有向超图结构代替平面命名实体识别中常用的无向图结构，通过设计标记模式来识别嵌套实体，但这种方法在设计超图时需要大量人工，而且当句子过长或实体类别过多时，超图结构会很复杂。另一种解决方法是利用片段抽取分类的方法，该方法通过抽取出句子中的子序列并进行分类来完成实体识别，但这种方法需要找出句子中所有的子序列，计算代价大，且在表示子序列时没有充分利用到文本特征。

针对上述问题，本文提出一种边界强化分类模型。将卷积神经网络（Convolutional Neural Network，CNN）和多头注意力机制应用到序列标记模型中，增强模型检测实体边界的能力。在进行分类时，结合实体片段关键字和实体边界词进行实体表示，从而更充分地利用模型学到的特征。在此基础上，引入多任务损失对模型进行训练，进一步优化模型性能。

1 相关工作

早期关于nested NER的研究主要采用手工特征的方法或是依赖规则的处理，如文献[5-6]采用隐马尔可夫模型识别内部的平面实体，然后通过基于规则的方法获取外部实体，文献[7]提出基于支持向量机的方法来提取嵌套实体。但是这些依赖手工特征和规则的方法移植性差、容错率低，需要消耗大量的人力和时间。

随着深度学习的不断发展以及NER应用日渐成熟，对于nested NER的研究也主要集中在了深度学习方法上。文献[8-9]利用超图解决nested NER问题，通过设计超图来表示所有可能的嵌套结构，然后从超图标签中恢复嵌套实体。但是为了避免虚假结构和结构歧义，在设计超图时需要消耗大量的人力，并且这种使用超图的方法在训练和推理中有着较高的时间复杂度。文献[10]提出一种多层神经网络模型，通过动态堆叠flat NER层来识别嵌套实体，其中flat NER层利用BiLSTM捕获序列的上下文表示，然后将其输入到CRF中。但这种方法由于下一层的计算依赖上一层的表示结果，因此无法并行计算，同时层级之间容易出现错误传递。而且由于嵌套层数较深的实体数量较少，因此该方法还存在深层实体稀疏的问题。文献[11]提出的锚区域检测模型，通过对实体的头部驱动短语结构进行建模来检测嵌套的实体，即先使用锚点词检测器检测句子中的每个词，判断其是否为实体的锚点词，并且判断其对应的实体类别，再设计区域识别器来识别以每个锚点词为中心的实体边界，最终完成嵌套实体的检测。文献[12]则通过将词或实体合并形成新实体来完成嵌套实体的检测。

对于nested NER任务，另一种解决方法是对文本进行片段抽取分类。文献[13]提出一种考虑所有可能片段的嵌套神经网络穷举模型，先穷举出所有小于最大实体长度的片段，再对每个片段进行向量表示，最后通过一个分类器为每个片段打上实体类型或非实体类型的标签。该模型虽然避免了层级检测模型错误传播的问题，但由于要对所有可能的片段进行分类筛选，因此计算开销大、推理效率低，而且由于缺乏精确的边界信息，抽取出的片段多为非实体片段。文献[14]提出使用一种局部探测方法，利用固定长度的片段及其前后文对实体进行识别。文献[15]提出一种边界感知的神经网络模型，将序列标注模型应用到片段抽取中，通过序列标注模型检测边界来精确定位实体，并在检测所得边界的基础上，对边界内的实体进行向量的平均表示。文献[16]结合词性进行实体边界检测，对实体边界检测和分类进行共同训练。虽然文献[15-16]模型利用了边界信息，但在表示实体时没有充分利用到实体特征信息。

综上可知，现有的嵌套命名实体识别方法存在对实体边界的检测不够准确、未有效利用实体边界信息的问题。本文提出一种边界强化分类模型进行嵌套实体的识别。利用不同卷积核大小的卷积神经网络获取邻近单词的特征，利用多头注意力强化模型的表达能力，从而获得有效的实体边界信息。同时，在表示实体时充分利用实体特征信息，计算实体内部各词权重并结合实体边界词，得到准确的实体表示。

2 边界强化分类模型

本文边界强化分类模型的整体结构如图1所示。该模型由以下3个部分组成：1）词表示层，用于提取字符之间的依赖信息，得到准确的单词表示；2）特征提取层，用于提取单词的局部信息和上下文信息；3）实体识

别层，用于对特征提取层输出的单词特征进行实体边界的分类，对类别为“B”和“E”的单词进行匹配，将匹配得到的区间作为候选实体，并对候选实体的类别进行预测。

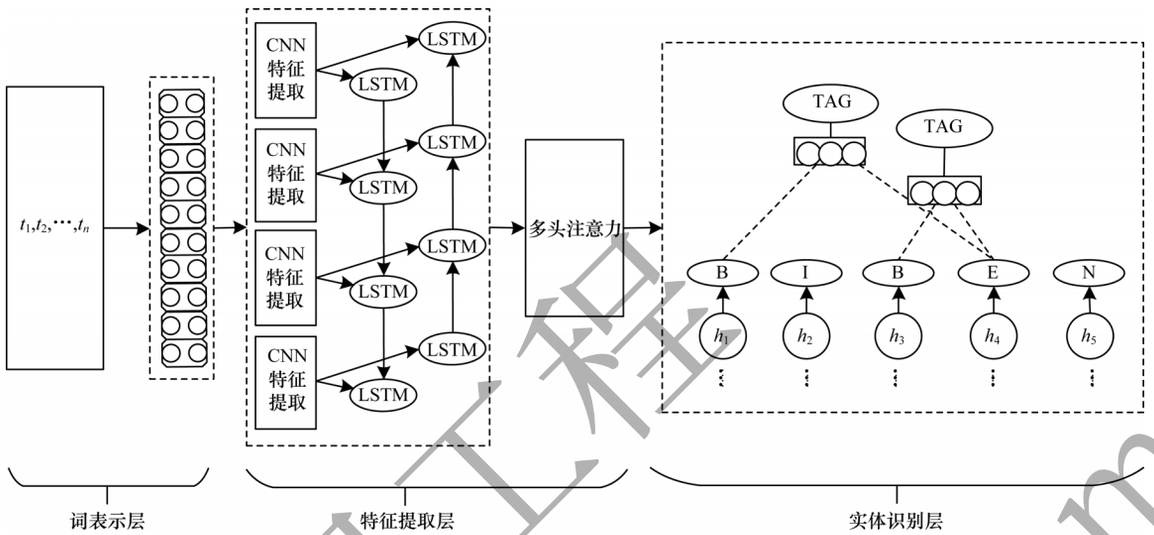

图1　边界强化分类模型架构

Fig.1　Framework of boundary enhanced classification model

2.1　词表示层

字符中包含单词的语法、语义等信息，提取字符中包含的特征已经在自然语言处理中广泛应用，如文献[17]在研究flat NER时将字符嵌入结合到词嵌入中，取得了较好的结果。因此，本文在进行词嵌入表示时将字符级别的特征与词级别的特征进行融合，作为最终单词的表示。

假设句子由n个词组成$[t_1,t_2,\cdots,t_n]$，通过查找预先训练好的词嵌入表得到每个词t_i对应的词嵌入w_i：

$$w_i=e^w(t_i) \tag{1}$$

其中：e^w为预训练得到的词嵌入表。对于词t_i中的每个字符c_i，对其进行随机初始化得到字符嵌入表示c_i'并输入到BiLSTM[18]中，以获取单词的拼写和形态特征，如式(2)所示：

$$w_{c_i}=\left[\overrightarrow{h}_{c_i};\ \overleftarrow{h}_{c_i}\right] \tag{2}$$

其中：$\overrightarrow{h}_{c_i}$和$\overleftarrow{h}_{c_i}$分别表示BiLSTM中正向和反向输出，将其拼接起来作为字符表示。最后，将w_i与w_{c_i}进行拼接，得到单词表示：

$$x_i=\left[w_i;\ w_{c_i}\right] \tag{3}$$

并将x_i作为最终的单词表示，输入到后序网络中。

2.2　特征提取层

长短期记忆(Long Short Term Memory Network, LSTM)网络虽然可以进行序列建模，但局部特征提取能力不如CNN，只依靠LSTM可以学习到上下文特征，但却忽略了局部的特征。因此，本文使用CNN-BiLSTM作为文本特征提取模型，将CNN关注局部信息特征和BiLSTM能够学习文本长距离依赖关系的优点相结合来获取文本语义特征。

2.2.1　CNN特征提取

卷积神经网络[19]在自然语言处理中多为一维卷积(Convolution1D)。因此，本文模型的卷积层使用一维卷积，卷积核宽度与输入的单词向量的嵌入维度相同，高度采用1、3、5这3种不同尺寸的卷积核提取词本身以及邻接词的信息，从而获得短距离单词间的依赖信息，为后续实体的表示提供更丰富的局部信息。

在本文模型中，第1层卷积分别使用100个高度为1、50个高度为1、50个高度为3以及100个高度为3的卷积核进行卷积，第2层卷积分别使用50个高度为3的卷积核对100个高度为1情况下的结果进行卷积，使用50个高度为5的卷积核对100个高度为3情况下的结果进行卷积。多尺度卷积的计算如式(4)所示：

$$z_i=f(w\cdot x+b) \tag{4}$$

其中：z_i表示卷积操作得到的第i个特征；w表示一个卷积核的权重；f表示非线性函数ReLU；b表示偏置。

通常在卷积层后会加池化层来提取特征，但在池化时会丢失部分细节特征。因此，本文在卷积后没有进行池化操作，而是保留卷积结果，将最终的4种卷积结果进行拼接，作为后续模型的输入，如式(5)所示：

$$g=[z_1';\ z_2;\ z_3;\ z_4'] \tag{5}$$

其中：z_i'表示第2层卷积后的结果。本文模型中的CNN特征提取过程如图2所示。

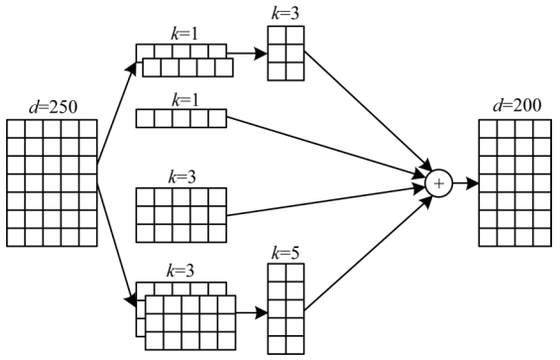

图2 多尺度CNN特征提取过程

Fig.2 Process of multi-scale CNN feature extraction

2.2.2 LSTM特征提取

本文模型使用BiLSTM作为特征提取器。单向的LSTM只能获得前文的信息，而BiLSTM能够从2个方向学习到序列中的上下文信息，更有效地将单词在上下文中的隐藏状态表示出来。BiLSTM中隐藏状态表示如式(6)~式(8)所示：

$$\overrightarrow{h_i} = \overrightarrow{\mathrm{LSTM}}(g_i, \overrightarrow{h_{i-1}}) \tag{6}$$

$$\overleftarrow{h_i} = \overleftarrow{\mathrm{LSTM}}(g_i, \overleftarrow{h_{i-1}}) \tag{7}$$

$$h_i = \left[\overrightarrow{h_i}; \overleftarrow{h_i}\right] \tag{8}$$

其中：g_i是上节中利用不同卷积核拼接的结果；$\overrightarrow{h_i}$和$\overleftarrow{h_i}$分别表示h_i在BiLSTM中向前和向后的隐藏状态。BiLSTM输出的结果$\boldsymbol{H}=\{h_1,h_2,\cdots,h_n\}$将被输入到后续模型中。

2.2.3 多头注意力机制

虽然BiLSTM能够学习语序信息，进而获取实体间的依赖关系，但在处理长句子时其效果会有所影响。而多头注意力机制通过多个子空间表示来提升模型关注不同特征的能力，这对后续实体的边界分类和类型分类都有所帮助。因此，本文在BiLSTM后加入多头注意力来提升模型性能。多头注意力Attention($\boldsymbol{K},\boldsymbol{Q},\boldsymbol{V}$)的计算公式如式(9)所示：

$$\mathrm{Attention}(\boldsymbol{K},\boldsymbol{Q},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{h}}}}\right)\boldsymbol{V} \tag{9}$$

其中：$\boldsymbol{K}$、$\boldsymbol{Q}$、$\boldsymbol{V}$分别表示查询矩阵、键矩阵和值矩阵；d_{h}表示嵌入维度；$\boldsymbol{H}=\{h_1,h_2,\cdots,h_n\}$表示之前BiLSTM的输出。本文设置$\boldsymbol{K}=\boldsymbol{Q}=\boldsymbol{V}=\boldsymbol{H}$。

通过投影到不同子空间来获得不同的注意力，并将不同子空间的特征进行连接，如式(10)、式(11)所示：

$$H^* = \mathrm{concat}(T_1; T_2; \cdots; T_k) \tag{10}$$

$$T_i = \mathrm{attention}(Q_i, K_i, V_i) \tag{11}$$

其中：Q_i、K_i、V_i是网络中学习的参数；k表示网络中设置的子空间个数；T_i表示输入的句子对第i个子空间提取的特征。

2.3 实体识别层

对于嵌套在内部的实体，难以确定实体层数和实体个数，采用多层的CRF模型会造成错误传播的问题。因此，本文利用片段抽取分类的方法，通过检测实体边界，先抽取出候选的实体片段(包括嵌套实体片段)，再进行标签类别的判断。

2.3.1 实体边界检测

虽然分层序列标注思想和穷举区域分类思想各有不足，但它们是相辅相成的。之前的工作对2种方法的使用过于分离，受文献[15]启发，本文在进行实体识别时，同样利用序列标注的思想来判断边界标签。将序列$T[t_a,t_{a+1},\cdots,t_b]$($a<b$)中的任一实体表示为$R_{i,j}$，表示该实体由序列$[t_i,t_{i+1},\cdots,t_j]$($a\leqslant i,j\leqslant b$)组成。在预测边界标签时，计算序列中每个词对应的实体边界标签，将实体的开始位置标记为‘B’标签，实体结束位置标记为‘E’标签，实体中间位置标记为‘I’标签，非实体标记为“N”。

对T中的每个词t_m，将与之对应的通过多头注意力得到的特征h_m^*输入到激活函数LeakyReLU中计算其可能的边界标签d_m，然后利用softmax分类器，得到该词最可能的边界标签d_m^*，如式(12)、式(13)所示：

$$d_m = \mathrm{LeakyReLU}(h_m^*) \tag{12}$$

$$d_m^* = \mathrm{softmax}(d_m) \tag{13}$$

其中：h_m^*为经过多头注意力后的输出特征。由于多数单词的标签为“N”，少数的标签为边界标签，因此边界分类器在训练过程中可能会遇到类别不平衡的问题。为了缓解这种类别不平衡现象对分类模型造成的影响，本文采用FocalLoss来计算损失，如式(14)所示：

$$l_{\mathrm{d}} = \sum -\alpha(1-\hat{d}_m^*)^{\gamma}\ln\hat{d}_m^*\gamma \tag{14}$$

其中：α为权重因子；γ为可调节因子。

对于序列$T[t_a,t_{a+1},\cdots,t_b]$，计算出$T$中各词对应的标签后，对所得边界标签序列$D[d_a^*,d_{a+1}^*,\cdots,d_b^*]$中所有边界标签为‘B’和‘E’的词进行匹配，将匹配得到的词片段作为候选实体。

2.3.2 标签类别判断

得到候选实体后对其进行分类，为目标实体分配正确标签，将非目标的候选实体排除。分类时首先要计算候选实体的向量表示。传统方法在对抽取出的实体进行表示时，多是采用实体片段的平均表示，即把实体片段内的词向量相加，再除以片段长度即词向量的个数，得到该实体的表示。但是这种方法没有考虑到实体片段内不同词对实体类型分类的贡献度，如词序列“the department of education”是一个标签应为“organization”的实体，在该实体中，词“department”应为关键词，其对实体最终标签分类所作的贡献应大于其他词。因此，在实体表示时，词“department”的表示在整个实体表示中所占比重应更大。此外，实体的边界词包含了实体片段的边界信息。为了充分利用这些信息，获得更准确的实体表示，本文在进行实体表示时将实体内关键字与实体头尾词相结合。实体边界检测和标签类别判断示意图如图3所示。

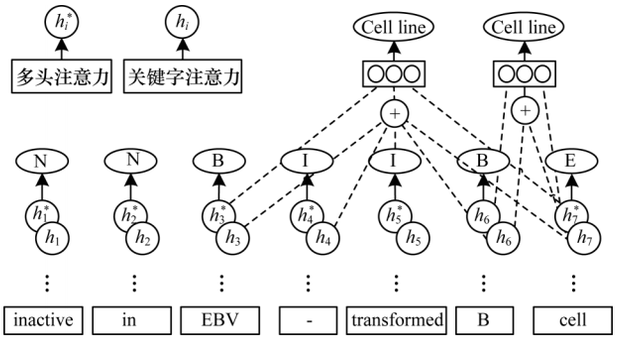

图3　实体边界检测和标签类别判断示意图

Fig.3　Schematic diagram of entity boundary detection and tag category judgment

本文使用前馈神经网络（Feedforward Neural Network，FFNN）进行实体片段关键字注意力的计算，如式（15）~式（18）所示：

$$\mu_m = \mathrm{FFNN}(h_m) \tag{15}$$

$$\hat{\mu}_m = \frac{\exp(\mu_m)}{\sum_{\mathrm{start}(i)}^{\mathrm{end}(j)} \exp(\mu_k)} \tag{16}$$

$$r_{i,j} = \sum_{\mathrm{srart}(i)}^{\mathrm{end}(j)} \hat{\mu}_m h_m \tag{17}$$

$$R_{i,j} = [h^*_{\mathrm{start}(i)}; h^*_{\mathrm{end}(j)}; r_{i,j}] \tag{18}$$

其中：h_m是第t个词经CNN-BiLSTM后得到的隐藏状态。在得到候选实体$R_{i,j}$的表示后，将其输入到ReLU激活函数中，并用softmax对结果进行分类，得到最终的实体标签，如式（19）所示：

$$d_{i,j} = \mathrm{softmax}(U_{i,j} R_{i,j} + b_{i,j}) \tag{19}$$

其中：$U_{i,j}$和$b_{i,j}$是网络中学习的参数。同样地，考虑到实体的类别数量不同，为缓解实体类别不平衡现象对实体分类模型造成的影响，使用FocalLoss对实体分类器进行优化，如式（20）所示：

$$l_{\mathrm{D}} = \sum -\alpha(1-\hat{d}_{i,j})^{\gamma} \ln \hat{d}_{i,j} \tag{20}$$

2.4　联合训练和推理

由于本文模型在进行实体边界检测和类别判断时共享相同的实体边界，因此对实体边界检测和实体类别判断这2个任务的损失进行联合训练。在训练阶段，将数据的真实实体边界标签输入到模型中来训练实体边界检测分类器，避免分类器在训练时受到错误边界检测影响。在测试阶段，边界检测分类器的输出用于指示在预测分类标签时应考虑哪些实体片段。

模型的总损失为实体边界检测和实体类别判断两部分的损失之和，如式（21）所示：

$$L = \lambda \mathrm{l}_{\mathrm{d}} + (1-\lambda) l_{\mathrm{D}} \tag{21}$$

其中：λ是一个超参数，用于控制实体边界检测和标签类别判断两部分损失的重要程度。

3　实验与分析

3.1　数据集介绍

在GENIA、GermEval 2014、JNLPBA这3个公共数据集上进行实验，评估本文模型性能。其中，GENIA和GermEval 2014是nested NER中常用的公共数据集。

GENIA数据集是基于语料库GENIA3.0.2构建而成的，共有语句18 546句。按照之前Finkel and Manning[20]以及文献[8]的划分规则，将其中的实体类型DNA、RNAs、protein subtypes分别归类到实体类型DNA、RNA、Protein中，保留Cell line和Cell type实体类型，除去其他实体类型。最终，数据集中包含5种实体类型：DNA，RNA，Protein，Cell line，Cell type，其中嵌套实体占比10%。数据集以8.1∶0.9∶1的比例分为训练集，验证集和测试集，即训练集包含语句15 023句，验证集包含1 669句，测试集包含1 854句。5种实体类型的具体数量如表1所示。

表1　GENIA数据集统计指标

Table 1　GENIA dataset statistics

实体类型	训练集	验证集	测试集	总计	嵌套数
DNA	7 650	1 026	1 257	9 933	1 744
RNA	692	132	109	933	407
Protein	28 728	2 303	3 066	34 097	1 902
Cell line	3 027	325	438	3 790	347
Cell type	5 832	551	604	6 987	389
总计	45 929	4 337	5 474	55 740	4 789

GermEval 2014数据集是一个德语语料的嵌套命名实体数据集，本文使用这个数据集来评估模型在不同语言中的性能。GermEval 2014数据集中含有31 000多条语句，对应有590 000多个词。

JNLPBA数据集来源于GENIA语料库，在该数据集中仅保留最外层的实体。因此，本文使用这个数据库来评估模型在识别flat NER方面的表现。在实验中，按照与GENIA数据集相同的划分设置将JNLPBA数据集中的实体子类别合并为5种类别。

3.2　实验参数设置

本文模型基于PyTorch框架实现，预训练的词向量维度为200维，字符嵌入的维度为50维并随机初始化。模型其他参数的取值如表2所示。

表2　模型参数设置

Table 2　Setting of model parameters

模型参数	设置
CNN卷积核大小	（1,3,5）
多头注意力头数	4
BiLSTM维度	200
Dropout比率	0.5
学习率	0.000 4
λ	0.3

3.3 评估方法

采用整体准确率、召回率和F1值(F1-score)作为最终的评价指标,且使用一个严格的评估标准,即当实体边界的分类和实体类别的分类同时正确时,才会认为实体分类正确。准确率、召回率和F1-score的计算公式如式(22)~式(24)所示:

$$P=\frac{T_P}{T_P+F_P} \tag{22}$$

$$R=\frac{T_P}{T_P+F_N} \tag{23}$$

$$F_1=\frac{2PR}{P+R} \tag{24}$$

其中:T_P表示模型正确识别出的真实体数目;F_P表示实际为假实体但模型识别为真实体的实体数目;F_N表示实际为真实体但模型并未识别出的实体数目。

3.4 对比模型

为验证本文模型的有效性与性能优势,选取与本文模型相关的且具有代表性的模型作为基线模型,具体如下:

1)文献[8]模型。该模型使用基于超图的方法来解决命名实体识别问题。

2)文献[21]模型。该模型对基于超图的方法进行了改进。

3)文献[13]提出的深度穷举模型。该模型通过穷举出所有可能片段来找出实体。

4)文献[10]提出的多层的CRF模型。该模型以由内向外的方式识别不同级别的嵌套实体。

5)文献[11]模型。该模型先识别锚点词,再识别以锚点词为中心的实体边界,进而确定实体。

6)文献[15]模型。该模型利用序列标注模型检测嵌套实体边界,合并相应的边界标签序列,完成分类预测。

7)文献[16]模型。该模型结合词性进行实体边界检测,对实体边界检测和分类进行共同训练。

3.5 实验结果与分析

不同模型在GENIA数据集上的性能指标如表3所示,其中,加粗数据表示最优值。由表3可知,本文模型在召回率和F1-score上都优于其他对比模型,准确率仅低于文献[16]模型,优于其他模型。

表3 不同模型在GENIA上的性能指标

Table 3 Performance index of different models on GENIA %

模型	准确率	召回率	F1-Score
文献[8]模型	72.5	65.2	68.7
文献[21]模型	75.4	66.8	70.8
文献[13]模型	73.3	68.3	70.7
文献[10]模型	76.1	66.8	71.7
文献[11]模型	75.2	73.3	74.2
文献[15]模型	75.9	73.6	74.7
文献[16]模型	**78.9**	72.7	75.7
本文模型	77.7	**74.5**	**76.1**

从整体上看,基于超图的方法及文献[13]模型准确率和召回率略低于其他模型,而文献[10]模型可能受到错误在分层模型中逐层传递的影响,其召回率也有所不足,而对边界分类和实体分类加以约束的方法在效果上有所提升。

相较于文献[15]模型,本文模型在准确率上获得1.8%的提升,在召回率上,本文模型达到最优的74.5%,这说明本文模型识别出的实体绝大多数是真实有效的,这是因为结合实体片段关键字和实体头尾词的方法为模型提供了更准确的实体表示,使得分类器有能力确定候选片段是否为有效的实体。因此,本文模型获得了更好的F1-Score,表明本文模型能够更准确地识别实体包括嵌套实体。

表4对比了不同模型在GENIA的5种实体类型检测上的性能表现,其中,加粗数据表示最优值。可以看出,文献[15]模型在RNA上F1-score略高于本文模型,而在其他不同类型的实体上,本文模型性能较对比模型有不同程度的提高。

表4 不同模型在GENIA上实体类型检测的性能指标

Table 4 Performance index of entity type detection of different models on GENIA %

实体类型	准确率	召回率	F1-Score			
			本文模型	文献[10]模型	文献[13]模型	文献[15]模型
DNA	75.9	69.2	**72.4**	70.1	67.8	70.6
RNA	85.0	78.0	81.3	80.8	75.9	**81.5**
Protein	78.9	76.8	**77.9**	72.7	72.9	76.4
Cell line	72.6	70.3	**71.5**	66.9	63.6	71.3
Cell type	78.0	69.4	**73.4**	71.3	69.8	72.5
总计	77.7	74.5	**76.1**	71.1	70.7	74.7

本文模型与文献[15]模型检测句子中每个词对应边界标签的性能对比如表5所示。可以看出，本文模型结果优于文献[15]模型的结果，说明加入多尺度CNN和多头注意力后模型能够更高效准确地判断边界标签的类型。

表5　不同模型在GENIA上实体边界标签预测的性能指标

Table 5　Performance index of entity type boundary tag prediction of different models on GENIA　%

边界标签	准确率	召回率	F1-Score	
			本文模型	文献[15]模型
N	99.3	99.1	99.2	99.2
B	85.5	84.4	84.9	84.3
I	83.2	89.9	86.3	85.6
E	86.3	87.9	87.1	86.6

在GermEval 2014数据集上的实验结果如表6所示，其中，加粗数据表示最优值。可以看出，本文模型在召回率与F_1-Score上较其他对比模型均有所提升。对比表3和表6可以发现，在GENIA数据集上的效果比GermEval 2014数据集上的效果好，这是因为GENIA数据集中嵌套实体数量更多，而GermEval 2014数据集中实体较为稀疏。

表6　不同模型在GermEval 2014上的性能指标

Table 6　Performance index of different models on GermEval 2014　%

模型	准确率	召回率	F1-Score
文献[13]模型	**75.0**	60.8	67.2
文献[10]模型	72.9	61.5	66.7
文献[15]模型	74.5	69.1	71.7
本文模型	74.6	**70.2**	**72.3**

此外，为验证本文模型在Flat NER上的适用性，在JNLPBA数据集上进行实验，实验结果表明，F1-Score达到74.2%。

3.6　消融实验

为验证CNN-BiLSTM特征提取、多头注意力机制以及结合关键字和实体边界词表示实体的有效性，设计以下5种模型进行对比，并在GENIA数据集上进行实验与分析。

1）BiLSTM+M-head-att+SEKey：只使用BiLSTM进行上下文特征提取，在BiLSTM之后使用多头注意力机制，实体分类时使用结合实体边界词和关键字的方法。

2）CNN-BiLSTM+SEKey：使用CNN和BiLSTM进行文本的信息特征提取，之后不使用多头自注意力机制，直接使用结合实体边界词和关键字的分类方法进行实体分类。

3）CNN-BiLSTM+M-head-att+Mean：使用CNN和BiLSTM进行文本的信息特征提取，之后使用多头注意力机制，实体分类时不使用结合实体边界词和关键字的方法，而是将实体片段内的词向量表示进行平均化，得到实体表示。

4）CNN-BiLSTM+M-head-att+Key：使用CNN和BiLSTM进行文本的信息特征提取，之后使用多头注意力机制，实体分类时不结合实体边界词，只使用关键字注意力的方法来得到实体表示。

5）CNN-BiLSTM+M-head-att+SEMean：使用CNN和BiLSTM进行文本的信息特征提取，之后使用多头注意力机制，实体分类时不使用关键字注意力的方法，使用结合实体头尾词和向量平均化表示的方法得到实体表示。

各个消融模型的实验结果如表7所示，与BiLSTM+M-head-att+SEKey相比，本文模型在特征提取时加入CNN有助于模型识别出真实有效的实体，这得益于不同的卷积核对长度不定的实体进行特征提取时起到了作用。相较CNN-BiLSTM+SEKey，使用多头注意力机制使得本文模型在准确率和召回率上都有所提高。CNN-BiLSTM+M-head-att+Mean和CNN-BiLSTM+M-head-att+Key的对比结果表明，单独使用关键字注意力方法来表示实体比向量平均化表示实体的整体性能值略高。CNN-BiLSTM+M-head-att+Mean、CNN-BiLSTM+M-head-att+SEMean以及CNN-BiLSTM+M-head-att+SEMean的对比结果表明，在表示实体时加上实体开始结束边界词后，模型的正确率都得到提升，且在召回率上提升尤为明显，说明在表示实体时加上实体边界词能够增强模型识别有效实体的能力。与CNN-BiLSTM+M-head-att+SE Mean的对比结果表明，本文模型准确率、召回率和F1-Score都有所提升，说明结合实体边界词和实体关键字的方法得到的实体表示更准确，且真实有效的实体数量更多，因此得到了更好的效果。

表7　消融模型在GENIA上的性能指标

Table 7　Performance index of ablation models on GENIA　%

模型	准确率	召回率	F1-Score
BiLSTM+M-head-att+SEKey	77.7	72.6	75.1
CNN-LSTM+SEKey	77.0	74.2	75.6
CNN-BiLSTM+M-head-att+Mean	78.3	72.4	75.3
CNN-BiLSTM+M-head-att+Key	77.3	73.5	75.4
CNN-BiLSTM+M-head-att+SEMean	76.8	74.4	75.7
本文模型	77.7	74.5	76.1

结合以上结果可知，综合使用CNN-BiLSTM、多头注意力机制进行特征提取以及结合关键字和实体头尾词来表示实体，能够有效提高模型的性能。

4 结束语

针对目前嵌套命名实体识别方法实体边界识别不准确、未能对特征信息有效利用的问题，本文提出一种注意力增强的边界强化分类模型。通过CNN-BiLSTM和多头注意力捕获邻近单词及上下文的依赖关系，获得更有效的文本特征，同时结合实体中的关键字和实体边界词的实体表示方法，强化实体的头尾信息和主要信息，为实体分类提供更准确的实体表示。后续将研究实体间的依赖关系以及词语与实体间的语义关系，进一步提升模型识别性能。

参考文献

[1] CHINCHOR N, ROBINSON P. MUC-7 named entity task definition [C]//Proceedings of the 7th Conference on Message Understanding. Washington D. C., USA: IEEE Press, 1995: 319-332.

[2] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. Berlin, Germany: Springer, 2001: 282-289.

[3] GUILLAUME L, MIGUEL B, SANDEEP S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2016: 260-270.

[4] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. [2021-05-12]. https://arxiv.org/abs/1508.01991.

[5] SHEN D, ZHANG J, ZHOU G, et al. Effective adaptation of a hidden Markov model-based named entity recognizer for biomedical domain[C]//Proceedings of the ACL 2003 Workshop on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2003: 49-56.

[6] ZHANG J, SHEN D, ZHOU G, et al. Enhancing HMM-based biomedical named entity recognition by studying special phenomena[J]. Journal of Biomedical Informatics, 2004, 37(6): 411-422.

[7] BORTHWICK A, GRISHMAN R. A maximum entropy approach to named entity recognition[D]. New York, USA: New York University, 1999.

[8] LU W, ROTH D. Joint mention extraction and classification with mention hypergraphs[C]//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2015: 857-867.

[9] WANG B, LU W. Neural segmental hypergraphs for overlapping mention recognition[C]//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2018: 204-214.

[10] JU M, MIWA M, ANANIADOU S. A neural layered model for nested named entity recognition[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: Association for Computational Linguistics, 2018: 1446-1459.

[11] LIN H, LU Y, HAN X, et al. Sequence-to-nuggets: nested entity mention detection via anchor-region networks[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2019: 5182-5192.

[12] FISHER J, VLACHOS A. Merge and label: a novel neural network architecture for nested NER[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2019: 5840-5850.

[13] SOHRAB M G, MIWA M. Deep exhaustive model for nested named entity recognition[C]//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2018: 2843-2849.

[14] XU M, JIANG H, WATCHARAWITTAYAKUL S. A local detection approach for named entity recognition and mention detection[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2017: 1237-1247.

[15] ZHENG C, CAI Y, XU J, et al. A boundary-aware neural model for nested named entity recognition[C]//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2019: 357-366.

[16] TAN C Q, QIU W, CHEN M S, et al. Boundary enhanced neural span classification for nested named entity recognition[C]//Proceedings of AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2020: 9016-9023.

[17] MA X Z, HOVY E. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2016: 1064-1074.

[18] WANG Y R, TIAN F. Recurrent residual learning for sequence classification[C]//Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2016: 938-943.

[19] 杨姗姗，姜丽芬，孙华志，等. 基于时间卷积网络的多项选择机器阅读理解[J]. 计算机工程，2020，46(11)：97-103.
YANG S S, JIANG L F, SUN H Z, et al. Multiple choice machine reading comprehension based on temporal convolutional network[J]. Computer Engineering, 2020, 46(11): 97-103. (in Chinese)

[20] JENNY R F, CHRISTOPHER D M. Nested named entity recognition [C]//Proceedings of 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2009: 141-150.

[21] MUIS A O, LU W. Labeling gaps between words: recognizing overlapping mentions with mention separators[C]//Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2017: 2608-2618.

编辑 金胡考