文章编号：1672-3961(2025)06-0035-10 DOI：10.6040/j.issn.1672-3961.0.2024.114

融合多特征和多头自注意力机制的高校学业命名实体识别

王禹鸥[1]，苑迎春[1,2*]，何振学[1]，何晨[1]
（1.河北农业大学信息科学与技术学院，河北 保定 071001；2.河北省农业大数据重点实验室（河北农业大学），河北 保定 071001）

摘要：为了有效解决高校学业领域实体归属和实体嵌套问题，提出一种基于多特征融合和多头自注意力机制的中文命名实体识别模型 Multi-feature BiGRU-MHSA-CRF（MBMC）。该模型从字、词、位置三个方面对文本语义特征进行表示，丰富多维度学业文本语义特征，并将特征向量输入到双向循环神经网络（Bi-directional recurrent neural network，BiGRU）以捕获全局语义特征。为了解决高校学业领域实体边界划分问题，对注意力机制进行改进，引入带有 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 权重矩阵的多头自注意力机制，增加学习参数，提升识别准确率，将所有可能的标签序列输出到条件随机场（conditional random fields，CRF），通过 CRF 解码生成实体标签序列。试验结果表明，该模型 F_1 值在公开数据集 CoNLL2003 和高校学业领域数据集分别达到 89.57%、86.14%，优于其它传统模型。

关键词：高校学业；命名实体识别；多特征融合；BiGRU；多头自注意力机制
中图分类号：TP391 **文献标志码**：A
引用格式：王禹鸥，苑迎春，何振学，等. 融合多特征和多头自注意力机制的高校学业命名实体识别[J]. 山东大学学报（工学版），2025，55(6)：35-44.
WANG Yuou，YUAN Yingchun，HE Zhenxue，et al. University academic named entity recognition based on the fusion of multi-feature and multi-head self-attention mechanism[J]. Journal of Shandong University（Engineering Science），2025，55(6)：35-44.

University academic named entity recognition based on the fusion of multi-feature and multi-head self-attention mechanism

WANG Yuou[1]，YUAN Yingchun[1,2*]，HE Zhenxue[1]，HE Chen[1]
（1. College ofInformation Science and Technology，Hebei Agricultural University，Baoding 071001，Hebei，China；2. Hebei Province Key Laboratory of Agricultural Big Data，Hebei Agricultural University，Baoding 071001，Hebei，China）

Abstract：In order to effectively solve the entity attribution and entity nesting in the academic domain of universities，a Chinese named entity recognition model Multi-feature BiGRU-MHSA-CRF（MBMC）was proposed based on multi-feature fusion. The text semantic features of the model were represented from three aspects of character，word and position to enrich the multi-dimensional semantic features of academic text. The feature vectors were fed into BiGRU（Bi-directional Recurrent Neural Network）to capture global semantic features. In order to solve the problem of entity boundary delimitation in the academic domain of higher education，the attention mechanism was improved by introducing a multi-head self-attention mechanism with $\boldsymbol{Q}$，$\boldsymbol{K}$，and $\boldsymbol{V}$ weight matrices and increasing the learning parameters to improve the recognition accuracy. All possible label sequences were output to the CRF，and the entity label sequence was generated by CRF decoding. The experimental results showned that the F_1 value of the model reached 89.57% and 86.14% in the public dataset CoNLL2003 and the college academic domain dataset，respectively. It was better than other traditional models.

收稿日期：2024-05-28
基金项目：国家自然科学基金资助项目（62102130）
第一作者简介：王禹鸥（2000— ），女，河北廊坊人，硕士研究生，主要研究方向为自然语言处理。E-mail：20222060106@ pgs.hebau.edu.cn
*通信作者简介：苑迎春（1970— ），女，河北保定人，教授，博士生导师，博士，主要研究方向为智能信息处理与大数据研究。
E-mail：nd_hd_yyc@ 163.com

Keywords: university academic; named entity recognition; Multi-feature fusion; Bi-directional recurrent neural network; Multi-head self-attention mechanism

0 引言

中国高等教育已迈入普及化阶段[1],在校生人数不断增多,进而学生存在的学业、生活问题呈现多样化、数量化趋势,如在校修读课程标准、考试规范、宿舍管理规定、毕业要求等[2]。这些问题的高效解答能有效帮助学生顺利完成学业并实现自身多元化发展。因此,如何快速准确地自动解答学生问题成为亟待解决的问题[3]。

精准识别学业领域实体是构建高校智能问答系统的关键技术之一[4],核心任务是抽取文本中具有特定意义的关键实体。然而高校学业领域文本中存在着大量学术专业词汇、课程名词缩写以及中英文夹杂的课程编号等各类实体,这对命名实体识别任务提出了挑战[5],主要体现在以下两个方面。

(1)实体归属问题:学业名词实体具有领域独特性,在不同语句中所做的成分不同。例如,"降级转专业后哪些公共课程可以申请免修?"中的三元组为〈公共课程,申请,免修〉,"降级转专业"不是三元组实体的组成部分,而在"降级转专业需要满足什么条件?"中的三元组为〈降级转专业,满足,条件〉,"降级转专业"是三元组实体的组成部分。二者虽为同一名词,但构成的三元组不同,即存在实体归属问题。

(2)实体嵌套问题:在高校学业领域内部分单独作为实体的名词同其他名词组合后能够构成与自身类别不同的实体。例如,在"每学期学生课程学分的上限是什么?"中,"课程学分"这一实体中包含"课程"和"学分"两个单独的实体,但是这句话强调的是"课程学分"这一整个实体,即存在实体嵌套问题。

近年来,深度学习方法[6]在命名实体识别领域得到广泛应用,如 BiLSTM 模型、BiGRU 模型等。文献[7]提出了一种结合 BiLSTM 和 CRF 的 BiLSTM-CRF 模型,效果优于简单的 CRF 和 LSTM 模型。目前,文献[8]构建 LF-BiLSTM-CRF 模型,提取中文药品不良反应的相关实体,协助构建中文药品不良评价体系。文献[9]针对文档级生物医学研究提出带有注意力机制的 BiLSTM-CRF 模型,确保文档中同一标记的多个实体的标记一致性。文献[10]针对食品安全领域研究提出 BiLSTM-Attention-CRF 模型,在 BiLSTM-CRF 中加入自注意力机制,以捕获对实体分类的显著性特征,提高实体分类准确率。文献[11]针对临床医学领域提出一种基于多任务注意力的 BiLSTM-CRF 模型,并且使用了预训练语言模型 ELMo,提高了识别准确率。文献[12]针对教育领域命名实体识别存在实体特征信息提取准确率低的问题,提出一种基于 BERT-BiLSTM-CRF 模型的命名实体识别研究方法,提高了教育领域命名实体识别准确率。文献[13]针对教育信息系统大量数据信息不能被正确识别的问题,提出一种结合 Bert-BiLSTM-Attention-CRF 的教育领域命名实体识别方法,确保识别教育信息数据的准确率。

由于 BiLSTM 结构复杂,收敛速度较慢,因此文献[14]提出了 BiGRU 模型,相较于 BiLSTM 模型,参数较少,收敛速度快,可以更灵活地控制信息流动,加速了迭代过程。

文献[15]针对教育领域实体边界识别不清晰的问题,在向量表示中融合字、词、位置信息,利用 BiGRU-CRF 进行序列建模,有利于更好地界定实体边界。文献[16]针对地质文本中存在大量长实体、嵌套实体问题,提出基于字词融合和注意力机制的 BiGRU 模型,提高实体识别准确率。文献[17]基于 BiGRU-CRF 模型将部首特征、字符特征、词特征进行有机结合,辅助 CNER 任务识别人名。

文献[18]提出基于多层次注意力机制的 BiGRU-CRF 模型,减少人工标注、错误标注问题,提高了识别准确率。文献[19]提出 Bert-BiGRU-CRF 模型,能有效提取上下文语义特征,在 BiGRU 层后添加了一层多头自注意力机制,缓解 BiGRU 提取局部特征能力的不足。文献[20]在 BiGRU-CRF 的基础上引入多头注意力机制,学习不同语义空间中远程实体与实体信息之间的依赖关系,有效提高了模型性能。但该语料库基于模式构建,相较于自动构建语料库性能较弱。

为解决上述实体归属和嵌套问题,本研究提出多特征融合[21]和多头自注意力机制模型的 MBMC 模型(Multi-feature BiGRU-MHSA-CRF)。本研究将多特征融合和注意力机制[22]纳入 BiGRU-CRF 模

型,通过多维度学习提取文本特征向量并输入 BiGRU,对输入的预训练文本向量进行深层次特征提取,捕捉输入序列中的上下文信息,实现特征学习;通过多头自注意力机制对文本序列的不同部分予以权重,学习不同实体的重要性以及它们之间的依赖关系,更有效地捕捉 BiGRU 神经网络输出信息;CRF 模型[23]通过定义状态转移矩阵实现约束关系,约束标注序列的合理性并可以学习到使得标注序列概率最大的参数,有效增强了对实体边界的辨识度,进而确保实体标注性能得到提高。

1　提出模型

该模型共分为三层:多特征融合层、BiGRU-MHSA 神经网络层和 CRF 标签约束层。模型架构如图 1 所示。图 1 中,$\boldsymbol{x}_i$ 为输入的文本向量,$\boldsymbol{W}_{x_i}$ 为本研究向量的字嵌入,$\boldsymbol{C}_{x_i}$ 为文本向量的词嵌入,$\boldsymbol{P}_{x_i}$ 为文本向量的位置嵌入,$\boldsymbol{D}_i$ 为经过循环神经网络处理后的向量,S_i 为注意力分数,$\boldsymbol{Y}_i$ 为向量最后得到的标签。

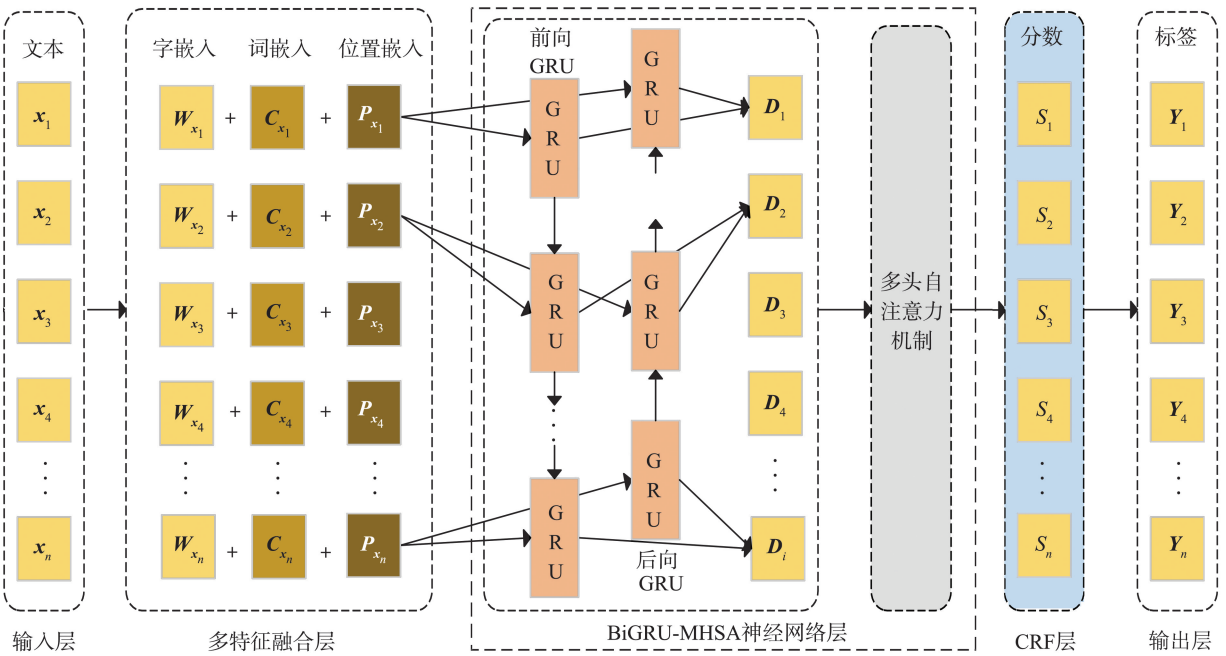

图 1　模型架构

Fig.1　Model framework

1.1　多特征融合层

本研究提出多特征融合的嵌入模型,添加对实体字符、词、位置信息的表示,多方位进行文本向量转化有助于模型更好地开展学习,使同一实体在不同语境下的向量表示不同,有效解决实体归属问题。

多特征融合层共从三个维度出发,分别是:字嵌入、词嵌入和位置嵌入。从汉字、词语和位置对文本语义特征进行融合表示。字嵌入采用 one-hot 编码将离散特征取值扩展到欧氏空间,相对特征之间的距离计算更为合理。通过该编码方式将学业领域文本中实体的每个字符转换到数值向量空间,构建向量矩阵。针对中文文本表述形式,词嵌入采用 jieba 分词工具对数据进行分词处理,获取字符对应的词,通过 Word2Vec 模型获取每个词的向量表示。从词语角度出发,避免字符之间的相互独立,将其形成关联更好的区分实体。该维度的向量转换在一定程度上增强了中文文本中词语的表示效果。位置嵌入在嵌入层加入了位置编码,表述输入的学业领域文本中每个 Token 的位置信息。本研究对不同位置的相同字符添加了不同编码,进而解决了字符语义表述不同的问题。将三个维度的嵌入层进行向量加和操作,得到多维度词嵌入矩阵 $\boldsymbol{B}=[\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_n]$,得到的嵌入矩阵输入到 BiGRU-MHSA 特征提取层。

1.2　BiGRU-MHSA 神经网络层

BiGRU-MHSA 特征提取模型主要由双向 GRU 层、多头自注意力机制和全连接层构成。该神经网络层利用 BiGRU 模型的双向性同时考虑文本序列过去和未来信息,对输入的预训练文本向量进行深层次特征提取,能够提高模型对嵌套实体的识别准确率。将 BiGRU 输出的向量由多头自注意力机制分配不同概率权重,最终由 CRF 层对前向和后向输

出向量完成拼接操作。

1.2.1 BiGRU 层

为了充分获取文本的上下文信息,捕获全局语义特征,本研究采用 BiGRU 模型对输入向量进行特征提取。模型结构如图 2 所示。BiGRU 包含重置门、更新门和两个隐状态,重置门和更新门的输出均通过 sigmoid 激活函数。重置门的作用是判断前一时刻所需要遗忘的隐状态信息,允许模型选择性地遗忘一些历史信息。更新门的作用是决定保留多少前一时刻的信息,用于对当前时刻的隐状态进行补充。

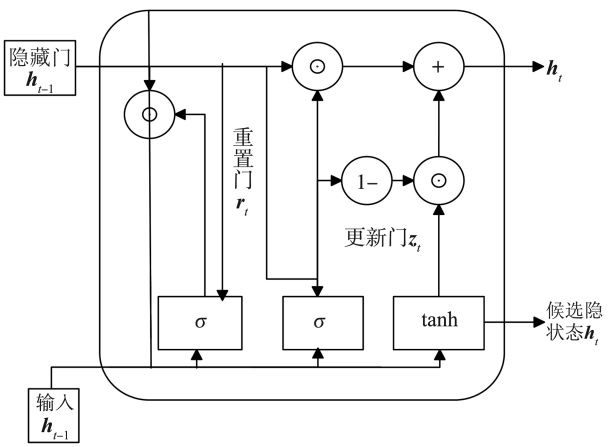

图 2 BiGRU 神经网络

Fig.2 BiGRU neural network

将句子第 t 个词语的词向量 $\boldsymbol{x}_t$ 作为 GRU 单元的输入

$$\boldsymbol{r}_t=\sigma(\boldsymbol{W}_r\boldsymbol{x}_t+\boldsymbol{W}_r\boldsymbol{h}_{t-1}+b_r), \tag{1}$$

式中,$\boldsymbol{r}_t$ 为重置门的输出向量,$\boldsymbol{W}_r$ 为重置门的权重矩阵,$\boldsymbol{h}_{t-1}$为前一时刻的隐状态,b_r 为重置门的偏置参数。

更新门的输出向量为

$$\boldsymbol{z}_t=\sigma(\boldsymbol{W}_z\boldsymbol{x}_t+\boldsymbol{W}_z\boldsymbol{h}_{t-1}+b_z), \tag{2}$$

式中,$\boldsymbol{W}_z$ 为更新门的权重矩阵,b_z 为更新门的偏置参数。

输出候选隐状态为

$$\tilde{\boldsymbol{h}}_t=\tanh\{\boldsymbol{W}\boldsymbol{x}_t+(\boldsymbol{r}_t\odot\boldsymbol{h}_{t-1})\boldsymbol{W}+b_{\tilde{h}}\}, \tag{3}$$

式中,$\boldsymbol{W}$ 为权重矩阵,$b_{\tilde{h}}$为候选隐状态的偏置参数。

$$\boldsymbol{h}_t=\boldsymbol{z}_t\odot\tilde{\boldsymbol{h}}+(1-\boldsymbol{z}_t)\odot\boldsymbol{h}_{t-1}, \tag{4}$$

式中,$\boldsymbol{h}_t$ 为当前时刻输出的隐状态。双向 GRU 由正向 GRU 和反向 GRU 组成,可以学习到前一时刻和后一时刻与当前状态之间的序列关系。通过双向 GRU 可以获取第 t 个词语的正向输出状态 $\vec{\boldsymbol{h}}_t$ 和反向输出状态 $\overleftarrow{\boldsymbol{h}}_t$,对两个方向的输出状态进行拼接,$\boldsymbol{h}_t=\vec{\boldsymbol{h}}_t+\overleftarrow{\boldsymbol{h}}_t$,构成 BiGRU 的输出状态,由此捕捉文本全局语义特征。

1.2.2 多头自注意力机制

在高校学业领域文本中,各实体重要程度不同,重要程度越高,实体包含的语义信息越关键。多头自注意力模块是由多个自注意力模块堆叠而来,可以根据重要程度,为学业领域文本动态分配权重。

与多头注意力机制相比,多头自注意力机制能够计算目标元素与其它所有元素(包括自身元素)之间的相似度,而不仅仅计算相邻两个元素之间的相似度,使模型可同时考虑上下文的多个方面,提升全局理解力。同时,多头自注意力机制可以捕捉 BiGRU 神经网络输出信息,有效解析文本间的联系,捕捉不同层次的结构信息,而后通过附加权重以获取词语间相似度,实现与词典的匹配并得到结构关系,从而使模型更好地区分嵌套实体边界。因此,本研究在特征提取中引入了多头自注意力机制,架构如图 3 所示,图中 h 为线形计算的次数。

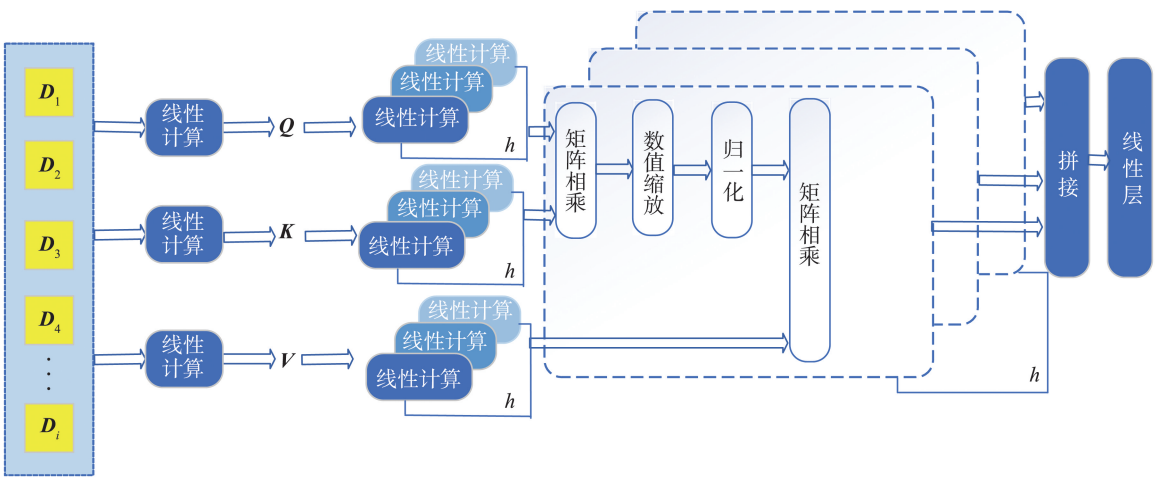

图 3 多头自注意力机制架构图

Fig.3 The structure of multi-head self-attention mechanism

自注意力机制仅通过语句内部进行相似度计算,无需借助外部标签的帮助,由相似度计算获取输入向量对标签分类的权重。多头自注意力机制即注意力机制运算进行 h 次,凭借多次计算学习到输入特征中的不同子空间所包含的相关信息。使用放缩点积的方法对于字符之间的相似度开展计算:

$$\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i=\boldsymbol{m}_i\cdot\boldsymbol{W}_Q,\boldsymbol{m}_i\cdot\boldsymbol{W}_K,\boldsymbol{m}_i\cdot\boldsymbol{W}_V, \tag{5}$$

式中:$\boldsymbol{m}_i$ 为输入序列;$\boldsymbol{Q}_i$、$\boldsymbol{K}_i$、$\boldsymbol{V}_i$ 分别为第 i 个元素的查询向量、关键向量和值向量,$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$ 为权重矩阵。

对于每一对元素 $\boldsymbol{m}_i$ 和 $\boldsymbol{m}_j$,都要计算一个注意力分数表示对的注意力程度:

$$S_i=\frac{\boldsymbol{Q}_i\boldsymbol{K}_j}{d_k}, \tag{6}$$

式中,S_i 为 $\boldsymbol{m}_i$ 对 $\boldsymbol{m}_j$ 的注意力程度,$\boldsymbol{Q}_i$ 为 $\boldsymbol{m}_i$ 的查询向量,$\boldsymbol{K}_j$ 为 x_j的值向量,d_k 为键向量的维度。

为避免向量内积过大导致较难进行归一化处理,所得值进行 Softmax 归一化处理后,实现权重分配。归一化如式(7)所示:

$$A_i=\mathrm{Softmax}(S_i)\cdot\boldsymbol{V}_i。 \tag{7}$$

通过不同的参数矩阵将 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$ 三个矩阵映射到多个不同平行的注意力头上开展多次注意力计算。每一个平行的注意力头对语句中不同位置的语义信息完成处理。

$$\boldsymbol{f}_i=\mathrm{Attention}(\boldsymbol{Q}_i\cdot\boldsymbol{W}_i,\boldsymbol{K}_i\cdot W_i,\boldsymbol{V}_i\cdot\boldsymbol{W}_i), \tag{8}$$

式中,$\boldsymbol{f}_i$ 为经过多次注意力机制运算后得到的矩阵。

最终,本研究对全部注意力头的运算结果实现拼接,得到包含更加丰富语义的文本向量矩阵,即句子新的表示。拼接计算方法为

$$M_i=W^M[e_1,e_2,\cdots,e_h], \tag{9}$$

式中,$\boldsymbol{M}_i$ 为经过线性变换得到的输出向量,W^M表示线性变换,e_h 为第 h 个注意力头的计算结果。

将拼接后的向量矩阵再次进行 Softmax 归一化处理,从而实现文本分类:

$$\boldsymbol{B}_i=\mathrm{Softmax}(\mathrm{Multi}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+b_1), \tag{10}$$

式中,$\boldsymbol{B}_i$ 为输出的预测标签,Multi 为多头注意力机制的函数表示。

1.3　CRF 标签约束层

经过 BiGRU 和多头自注意力机制层,文本中每个字符的各个标签分数已经被计算并输出,并且将其获得分数最大的标签作为输出标签。鉴于深度学习模型的训练效果无法判定标签得分的准确率,因此可能存在标签位置错误和标注乱序等问题。例如“实验课和实践类课程以及体育课能参加缓考吗?”中共包含四个实体,分别为课程实体“实验课”、“实践课”和“体育课”以及考试实体“缓考”。由于句中包含实体过多,需要过多连接词,因此实体标注的边界较易出现错误。本研究通过 CRF 模型对预测标签提供了约束,避免上述问题的出现,架构如图 4 所示。

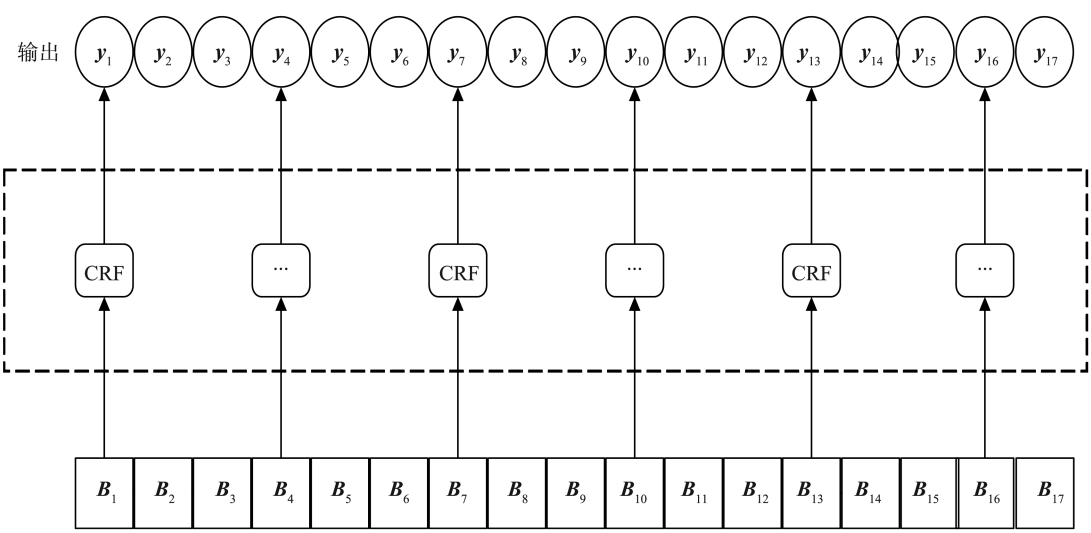

图 4　CRF 架构图

Fig.4　The structure of CRF

由图 4 可知,CRF 模型通过使用维特比算法来计算给定输入序列的最优标签序列,即具有最大概率的标签序列。利用上文输出的标签概率分布,即输入序列 $B=(B_1,B_2,B_3\cdots B_n)$,通过定义转移概率和发射概

率来计算输出给定标签序列的概率 $Y=(y_1,y_2,\cdots,y_n)$。定义 CRF 评估分数为

$$s_i=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i},\qquad(11)$$

式中,P_{i,y_i}表示第 i 个位置标签为 y_i 的分数,$A_{y_i,y_{i+1}}$为转移矩阵。

2 试验结果

2.1 数据集

2.1.1 数据集构建与标注

本研究在文献[24]的基础上,取自《中华人民共和国高等教育法》《国家教育考试违规处理办法》《普通高等学校学生管理规定》以及《河北农业大学学生管理手册》,通过归纳整理共计 150 000 余字作为试验数据集构建来源。命名实体识别任务中实体的设计原则是准确表达关键信息的语义。为更好地实现原始文本关键信息提取,对相关学业问题文本精准定位,本研究将实体类型分为十二类,分别为学分(CRE)、课程(C)、考试(EXAM)、学籍(STAT)、专业(MJR)、毕业(GRAD)、资助(SUBS)、奖惩(RAP)、宿舍(DORM)、学生社团(SO)、学生干部(SC)和其它(ELSE)。实体类型如表 1 所示。

表 1 实体类型

Table 1 Entity type

实体类型	举例
CRE	平均学分绩点
C	选课
EXAM	补考
STAT	入学手续
MJR	转专业
GRAD	毕业审核
SUBS	国家助学金
RAP	违纪行为
DORM	宿舍安全
SO	社团活动
SC	学生干部
ELSE	教务系统

2.1.2 数据集组成

完成上述预处理操作后,对试验文本原始数据完成整理筛选。参照公开数据集将样本划分为训练集、验证集和测试集,划分比例为 6∶2∶2。数据集中每一类别实体所占比重如图 5 所示。

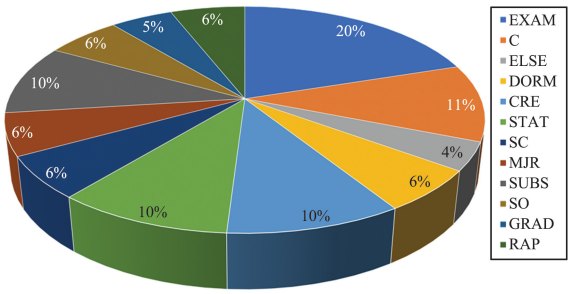

图 5 实体每一类型占比

Fig.5 Proportion of each type of entity

2.1.3 数据预处理

将有关规章制度以及河北农业大学教学管理文件和学生管理文件的文本数据分类,完成文本分割,并以单条语句为单位个体实现划分。本研究选用 CoNLL2003 公开数据集格式作为构建标准,将高校学生学业领域文本原始数据处理为字符标签占位符(character label placeholder, character LPH)的形式。其中,character LPH 为标签占位符,即在本研究数据中每一个字符与 character LPH 之间均存在空格。对原始中文文本数据进行切片处理后,根据所设计的文本分类规则,对实体信息实现预处理。

2.2 标注规则

命名实体标注策略共包含三种,分别为:三元标注策略 BIO、四元标注策略 BMES 和五元标注策略 BIOES。BIO 标注策略中 B 代表实体的第一个字,I 表示实体的剩余部分,O 表示非实体;BMES 标注策略中 B 表示实体的第一个字,M 表示实体的中间部分,E 表示实体的最后一个字,S 表示单独字词;BIOES 标注策略中 B 表示实体的第一个字,I 表示实体的中间部分,E 表示实体的最后一个字,S 表示单独成词的字符,O 表示非实体。

经对该研究领域文本归纳整理后分析,本研究所研究的高校学生学业领域文本数据中不存在字符实体,因此该语料整体采用 BIO 标注策略完成标注工作。以“EXAM”为例,相关实体的第一个字以“B-EXAM”标注,中间部分以“I-EXAM”标注,其余部分为“O”。高校学业领域文本分类试验语料标签设置如表 2 所示。

表 2 文本分类试验语料标签设置

Table 2 Corpus label setting for text classification experiment

分类标注样例																
对	考	试	成	绩	有	异	议	是	否	可	以	申	请	复	核	?
O	B-EXAM	I-EXAM	I-EXAM	I-EXAM	I-EXAM	I-EXAM	I-EXAM	O	O	O	O	O	O	O	O	O

2.3　评价指标

本研究命名实体识别研究选用了三类评价指标对实体识别效果进行评价。分别为:精确率(Precision, P)、召回率(Recall, R)和 F_1(F_1-Score)值。精确率为所有被识别的学业实体中被正确识别的实体概率,召回率为已标注的学业实体中被正确识别的实体概率,F_1 值为准确率与召回率的调和平均数。

$$P=\frac{T_P}{T_P+F_P}\times 100\%, \tag{12}$$

$$R=\frac{T_P}{T_P+F_N}\times 100\%, \tag{13}$$

$$F_1=\frac{2\times P\times R}{P+R}\times 100\%, \tag{14}$$

式中:T_P 表示模型预测为实体、实际也为实体的个数;F_P 表示模型预测为实体、实际不是实体的个数;F_N 表示模型预测不是实体、实际是实体的个数。

2.4　试验环境

本次中文命名实体识别研究具体环境如表 3 所示。

表 3　试验环境
Table 3　Experiment environment

环境	配置
OS	Windows 11
CPU	i7-10700
GPU	NVIDIA GeForce RTX 3060 12 GB
RAM	16 GB
TensorFlow	2.6.0
Python	3.8.8

2.5　结果与分析

2.5.1　MBCM 模型在 CoNLL2003 数据集下的性能验证试验

为了验证 MBMC 模型对高校学业领域命名实体识别的有效性,本节选用了以下模型进行比较,以下模型均为命名实体识别领域的经典模型。CoNLL 2003 数据集为命名实体识别领域通用数据集,包含了新闻文本中的实体类别和实体位置信息。其中,实体类别包括人名、地名、组织名和其他实体。

试验结果如表 4 所示。

(1) BiGRU:循环网络模型,由前向和后向两个独立的 GRU 单元构成,能够同时捕捉序列数据的前向和后向信息。

(2) BiGRU-Attention:循环神经网络-注意力机制模型,在 BiGRU 的基础上添加了注意力机制,允许模型对输入序列中的元素分配不同的权重,以便在处理数据时关注重要部分。

(3) BiGRU-CRF:在 BiGRU 的基础上添加了条件随机场,条件随机场通过计算局部概率和联合概率实现标签的预测和判断,处理标签之间的依赖关系。

(4) BiGRU-Multi_feature:在 BiGRU 的基础上添加多特征融合,多特征融合是指融合序列的多个属性特征,从而更好地学习文本的特征信息。

(5) BiGRU-Attention-CRF:结合了 BiGRU、注意力机制和条件随机场。

表 4　消融试验结果
Table 4　Results of ablation experiment

模型	P	R	F_1
BiGRU	0.802 8	0.732 3	0.765 9
BiGRU-Att	0.838 4	0.766 4	0.800 8
BiGRU-CRF	0.840 0	0.760 1	0.798 1
BiGRU-Multi_feature	0.853 2	0.765 8	0.807 1
BiGRU-Att-CRF	0.866 6	0.776 2	0.818 9
MBMC	**0.902 7**	**0.888 9**	**0.895 7**

注:加粗部分为该试验最优结果。

由表 4 可知,单一 BiGRU 模型在进行特征学习时更加注重全局特征提取。在结合注意力机制后,文本中的谓语作为可识别目标,突出待识别实体,同时减少其它噪声干扰。BiGRU-Attention 模型加强了对底层模型输出结果的捕捉能力,对于文本中所包含关键信息的提取性能有一定的提升,但是难以自动提取高阶组合特征中所包含的实体。CRF 模型的加入为整体实体识别及分类标注工作提供了标签约束作用,该层通过给定文本输入的正确标签序列,对完整序列进行标签的判定与预测。进而 BiGRU-Attention-CRF 模型对整体命名实体识别工作具有较好的抽取效果。

本研究提出的 Multi-feature BiGRU-MHSA-CRF 模型的三类评价指标分别达到了 0.902 7、0.888 9、0.895 7。与 BiGRU-Attention-CRF 模型在实体识别精准率上提升约 4%,综合评价指标提升约 8%。

由此说明,在面向中文文本命名实体识别任务中,模型应充分考虑其嵌入时所包含的信息,针对中文文本具有的特性更好地将词语和语义输入到模型,进而更好地开展实体识别。

2.5.2 句子级命名实体识别对比试验

开展了机器学习模型、深度学习模型与本研究提出模型之间的对比试验，结果如表5所示。

表5 句子级别下对比试验结果

Table 5 Comparative experimental results at the sentence level

模型	P	R	F_1
HMM	0.784 4	0.781 1	0.779 1
CRF	0.798 1	0.796 5	0.792 3
BiLSTM-CRF	0.824 0	0.820 7	0.822 3
BiGRU-CRF	0.852 9	0.828 1	0.840 3
BERT-BiGRU-CRF	0.860 4	0.828 8	0.844 3
BERT-BiLSTM-CRF	0.866 3	0.829 9	0.847 7
BERT-BiLSTM-Attention-CRF	0.879 0	0.835 2	0.856 6
MBMC	**0.886 5**	**0.837 6**	**0.861 4**

注：加粗部分为该试验最优结果。

（1）HMM：隐马尔科夫模型是机器学习模型，该模型以文本序列数据为输入，该序列对应的隐含信息为输出。

（2）CRF：条件随机场属于机器学习模型，条件随机场通过计算局部概率和联合概率来实现标签的预测和判断，处理标签之间的依赖关系。

（3）BiLSTM-CRF：在CRF基础上添加BiLSTM神经网络模型，BiLSTM由前向和后向两个独立的LSTM构成，能够同时捕捉序列数据的前向和后向信息。

（4）BERT-BiGRU-CRF：在BiGRU-CRF的基础上添加了Bert预训练模型，将文本转换成向量进行处理。

（5）BERT-BiLSTM-CRF：针对教育领域命名实体识别存在提取准确率低的问题，提出该命名实体识别方法。

（6）BERT-BiLSTM-Attention-CRF：在BERT-BiLSTM-CRF的基础上添加注意力机制。

以上各个模型指标均为进行了十折交叉验证试验后所得结果的平均值。本研究所提出的Multi-feature BiGRU-MHSA-CRF模型，在未分类标注下的三项评价指标分别达到0.886 5、0.837 6、0.861 4，均为最优。因为MBMC模型改变了原有词向量转换模式，从字、词、位置三个方面对文本语义特征进行表示，丰富多维度学业文本语义特征，对输入的预训练文本向量进行深层次特征提取，并引入了多头自注意力机制，文本向量的表述形式更丰富，针对中文文本特性加强了信息传递。同时，利用多头自注意力机制更有效地捕捉BiGRU神经网络输出信息，进而提升模型开展命名实体识别任务的能力。

HMM模型在该领域内命名实体识别效果较差，F_1值仅达到77.91%。因为HMM模型存在两个假设，它假设输出的观测序列之间是相互连续的并且每一次状态转移时仅与前一时刻的状态有关，无法有效衔接每一时刻状态。CRF模型在HMM基础上增加了全局概率。因此，CRF模型与HMM相比命名实体识别性能得到了提升。深度学习模型BiLSTM利用双向长短期特征提取优势，增强捕获全局上下文的能力，但在文本嵌入缺乏综合向量表示，进而影响了整体识别能力。BiGRU模型与BiLSTM模型相比，参数较少，门控单元结构简单，整体效果优于BiLSTM模型。BERT预训练模型本质上是一个个Transformer编码器，相比传统RNN模型更加高效，可以并行化处理同时能够捕捉长距离语义关系，但BERT预训练模型过于庞大，参数太多，且每个batch只有15%的token能够被预测，缺乏细粒度语义表示，相较于BiGRU-CRF模型提升效果不明显。BERT-BiLSTM-CRF模型利用BERT预训练模型获取输入序列语义的词向量，将训练好的词向量输入到BiLSTM模型中获取上下文特征，再根据CRF的标注规则和序列解码能力输出最大概率的序列标注结果。试验结果较HMM、CRF、BiLSTM-CRF、BiGRU-CRF均有提升。BERT-BiLSTM-Attention-CRF在BERT-BiLSTM-CRF的基础上添加了注意力机制，通过分配不同的权重，使模型更关注与当前任务相关的信息。但以上模型没有融合实体的字、词、位置信息，无法学习到实体深层次的语义特征，也没有采用多头自注意力机制学习实体包含的不同语义信息，无法更好地区分实体边界。本研究提出的MBMC模型既融合了实体的多特征，又采用了多头自注意力机制去识别实体边界，因此准确率与上述其他模型相比，达到最高。

2.5.3 分类标注下实体级命名实体识别对比试验

本研究结合河北农业大学的规章制度对高校学生学业领域文本进行分类标注，在各类别命名实体识别模型上完成对比试验研究工作，结果如表6所示。

表 6　分类标注下实体级识别结果

Table 6　Entity-level recognition results under classification and labelling

模型	F_1											
	C	EXAM	CRE	STAT	MJR	GRAD	RAP	DORM	SC	SO	SUBS	ELSE
HMM	0.693 0	0.550 5	0.818 2	0.670 7	0.840 0	0.767 4	0.728 1	0.800 0	0.690 8	0.706 5	0.760 7	0.708 3
CRF	0.728 1	0.786 1	0.857 1	0.701 4	0.809 6	0.783 0	0.841 6	0.910 7	0.847 1	0.788 5	0.823 1	0.753 7
BiLSTM-CRF	0.804 6	0.890 2	0.897 3	0.862 0	0.868 5	0.915 7	0.939 0	0.929 1	0.858 2	0.813 1	0.827 0	0.895 3
BiGRU-CRF	0.839 5	0.903 6	0.913 5	0.887 2	0.904 7	0.912 3	0.946 1	0.933 8	0.899 2	0.837 4	0.852 9	0.902 8
BERT-BiLSTM-CRF	0.846 1	0.905 8	0.917 4	0.887 5	0.913 4	0.913 5	0.946 4	0.934 6	0.906 2	0.850 2	0.869 5	0.907 1
BERT-BiGRU-CRF	0.857 3	0.909 9	0.920 5	0.887 8	0.926 6	0.912 3	0.946 5	0.935 8	0.918 4	0.863 7	0.883 0	0.912 4
Bert-BiLSTM-Attention-CRF	0.897 2	0.912 9	0.926 9	0.894 5	0.954 8	0.915 5	0.948 9	0.940 5	0.926 8	0.893 1	0.919 9	0.917 7
MBMC	**0.920 6**	**0.915 2**	**0.931 7**	**0.899 5**	**0.962 1**	**0.917 7**	**0.949 1**	**0.945 4**	**0.927 9**	**0.915 3**	**0.930 0**	**0.920 3**

注:加粗部分为该实验最优结果。

从表 6 可以看出,MBMC 模型以下特点。

(1) 综合评价指标 F_1 除了 STAT 类别外均大于 90%,相较于未进行分类时提升了大约 4%。在深度学习下命名实体识别任务已经被转换为多分类问题,对垂直领域完成分类标注工作后再开展的命名实体识别可有效针对中文文本不同词语开展实体识别工作。本研究提出的 MBMC 模型通过多特征融合增强鲁棒性,以三个维度的特征提取本研究领域文本中所包含的不同信息,将特征有效合理地组合后,模型整体分类性能增强,因此相较未分类下高校学业领域命名实体识别有进一步提升。

(2) MJR 类别 F_1 值高达 96.21%。因为该类别下所包含实体具有较强的规则性,且大多均包含“专业”两个中文汉字。例如“转专业的流程是什么?”“转专业学生,前后专业学费和学分收费标准不同怎么办?”等。双向门控循环神经网络有效利用前向和后向两个方向的历史信息,拼接实体所对应状态,提升模型学习能力。对于具有特定格式的实体,数据的表征能力和预测性能更优。对于该类实体边界,尤其是 I-MJR 识别效果更加。

(3) 在类别 C、EXAM 和 GRAD 中,本研究所整理的文本大多是该类别下的专有名词。例如:EXAM 中大多涉及到考试相关管理规定和违纪行为规范等,并且多数的文本中不会直接出现“考试”这一实体,而是以“通过伪造证件获得成绩会不会被认定为违纪?”等形式出现,该类待识别文本中包含较多实体且存在大量干扰信息。GRAD 中大多涉及到毕业要求等,实体并非直接指向毕业,例如“学生的学历学位授予规定是什么?”需要根据文本语义信息将“学位学历授予”判别为此类别(学历学位是毕业的必要条件)。因此,对这些类别的命名实体上边界和下边界的识别与其他标准实体相比较低。

(4) C 类别里存在选课部分,MJR 类别里存在转专业问题。转专业和选课大部分情况下是同时存在的,转专业学生都需要重新选课。因此这两类实体会同时出现在文本中,部分实体可能还会出现边界划分不明。对此类问题应该具体分析,判断嵌套实体在语义层面应该如何标注。BiGRU-CRF 模型能够学习词语的上下文信息,更好地捕获字词之间的依赖关系。对比 HMM、CRF 模型,MBMC 模型中 C 类别 F_1 值分别提升了 22.76%、19.25%,达到 92.06%。

3　结束语

本研究针对高校学生学业领域的现存问题,根据国家以及高校制定的相关管理规定文件和日常收集到的学生频发问题,构造适用于高校学业领域命名实体识别通用数据集,进而选择了合适的标注策略完成数据集标注工作。针对该领域内文本特性,提出 Multi-feature BiGRU-MHSA-CRF 模型,并加入了多头自注意力机制,有效解决了实体存在的实体归属和实体嵌套问题,提升了特有领域内命名实体识别的准确率。

该模型通过字符级别、词级别和位置级别三个维度,对原始数据文本开展了向量转化,在 Embedding 时丰富了所包含的信息,并通过深度学习神经网络 BiGRU,利用前向以及后向的 GRU 模型提升了对全局上下文以及局部特征的学习能力。利用多头自注意力机制对神经网络模型 BiGRU 输出结果进行有效捕获,最终利用 CRF 模型,对识别标签提供相应约束作用。该模型在本研究所构建的通用数据集上可有效开展命名实体识别任务研

究工作并取得了较好的效果，在句子级标注下 P、R、F_1 分别达到：0.886 5、0.837 6 和 0.861 4。在提出的分类标注开展命名实体识别工作时，单一类别识别综合评价指标 F_1 值更高达 96.21%。

本研究数据集标注范围仅针对学生日常所产生的问题和累积的学业问题，未涉及学业领域细小问题，还需不断细化完善该工作。在以后的工作中，可以尝试不同神经网络相结合的方法，继续深入开展研究。

参考文献：

[1] ZHAO Y, ZHANG A Y, LI X C, et al. RETRACTED: construction of a performance evaluation system for private higher education institutions in China based on balanced scorecard[J]. International Journal of Electrical Engineering & Education, 2023, 60(Suppl.1): 923-930.

[2] 孟平贵. 普通高等学校学生管理规定与学生工作规范化实务手册[M]. 长春：吉林音像出版社, 2005.

[3] FIGUEROA A. Automatically generating effective search queries directly from community question-answering questions for finding related questions[J]. Expert Systems with Applications, 2017, 77: 11-19.

[4] YANG Z J, WANG Y, GAN J H, et al. Design and research of intelligent question-answering (Q&A) system based on high school course knowledge graph[J]. Mobile Networks and Applications, 2021, 26(5): 1884-1890.

[5] LI J, SUN A X, HAN J L, et al. A survey on deep learning for named entity recognition [J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50-70.

[6] YADAV V, BETHARD S. A survey on recent advances in named entity recognition from deep learning models [C]// International Conference on Computational Linguistics, Santa Fe: USA, 2018: 20-26.

[7] GOYAL A, GUPTA V, KUMAR M. Recent named entity recognition and classification techniques: a systematic review[J]. Computer Science Review, 2018, 29: 21-43.

[8] 储德平, 万波, 李红, 等. 基于 ELMO-CNN-BiLSTM-CRF 模型的地质实体识别[J]. 地球科学, 2021, 46: 3039-3048.
CHU Deping, WAN Bo, LI Hong, et al. Geological entity recognition based on ELMO-CNN-BiLSTM-CRF model[J]. Editorial Committee of Earth Science-Journal of China University of Geosciences, 2021, 46: 3039-3048.

[9] LUO L, YANG Z H, YANG P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics, 2018, 34(8): 1381-1388.

[10] YUAN T P, QIN X Z, WEI C J. A Chinese named entity recognition method based on ERNIE-BiLSTM-CRF for food safety domain[J]. Applied Sciences, 2023, 13(5): 2849.

[11] YANG J L, LIU Y N, QIAN M H, et al. Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding[J]. Applied Sciences, 2019, 9(18): 3658.

[12] CHENG X W. Named entity recognition in the education domain based on BERT-BiLSTM-CRF-using data structures as an example[C]//2023 International Conference on Educational Knowledge and Informatization (EKI). 2023, Guangzhou, China: IEEE, 2023: 5-9.

[13] ZHANG J, JING S Y, HU J H. Named entity recognition method based on bert-bilstm-attention-CRF for education field [C]//2023 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). Hangzhou, China: IEEE, 2023: 383-388.

[14] KYUNGHYUN C, BARTVAN M, CAGLAR G, et al. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation[J]. Statistics, 2014: 1724-1734.

[15] 张召武, 徐彬, 高克宁, 等. 面向教育领域的基于 SVR-BiGRU-CRF 中文命名实体识别方法[J]. 中文信息学报, 2022, 36(7): 114-122.
ZHANG Zhaowu, XU Bin, GAO Kening, et al. SVR-BiGRU-CRF based Chinese named entity recognition for education domain[J]. Journal of Chinese Information Processing, 2022, 36(7): 114-122.

[16] 谢雪景, 谢忠, 马凯, 等. 结合 BERT 与 BiGRU-Attention-CRF 模型的地质命名实体识别[J]. 地质通报, 2023, 42(5): 846-855.
XIE Xuejing, XIE Zhong, MA Kai, et al. Geological named entity recognition combined BERT and BiGRU-Attention-CRF model[J]. Geological Bulletin of China, 2023, 42(5): 846-855.

[17] XU C W, WANG F Y, HAN J L, et al. Exploiting multiple embeddings for Chinese named entity recognition [C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China: ACM, 2019: 2269-2272.

[18] 李浩, 刘永坚, 解庆, 等. 基于多层次注意力机制的远程监督关系抽取模型[J]. 计算机科学, 2019, 46(10): 252-257.

（下转第 57 页）