融合领域词向量的实体识别研究

侯　敏，高　茂，张丽萍，闫　盛，赵宇博

(内蒙古师范大学 计算机科学技术学院，内蒙古 呼和浩特 010022)

摘　要：以字为切分单位的 BERT 预训练模型在实体识别任务中表现优异，但其忽略粗粒度的领域词汇作为整体的语义表示信息，对于教育领域课程文本中存在大量嵌套实体的识别效果不佳。针对上述问题，提出动态融合字、词级别词向量的 LEBERT-CRF 教育领域课程文本实体识别方法，利用词典适配器将领域词典特征高效融入 BERT 模型中，以提升 BERT 模型对实体词边界的识别效果，更好地适应课程知识抽取任务。结果表明，LEBERT-CRF 模型相较其他主流的实体识别模型表现更好，$F1$ 达到 95.47%。

关键词：实体识别；LEBERT；领域词向量；字词融合

中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1001－8735(2024)02－0197－10

DOI:10.3969/j.issn.1001－8735.2024.02.011

实体识别是自然语言处理任务中的一项基础研究，旨在识别非结构化文本中的特定实体名称及其对应的实体类型[1]。教育领域课程文本的实体识别是指从获取到的学科数据中，将与知识主题相关的概念、公式以及原理等提取出来，进而为知识抽取、实体链接以及智能化教育应用的研发与构建提供支撑，具有较大的研究价值[2]。

在实体识别任务中，BILSTM-CRF 因其在众多公开的数据集上良好的实验效果，成为深度学习技术的代表模型[3-4]。随着预训练模型的提出，多数研究者采用 BERT 结合 BILSTM-CRF 的方式进行实体识别[5]。谢腾等[6]针对传统的实体识别模型采用静态词向量导致无法处理“一词多义”的问题，利用 BERT 强大的上下文学习能力获取优质的词向量表示信息，在此基础上，将词向量输入至下游的 BILSTM-CRF 分类任务中进行实体识别，取得较好的准确率。但 BERT 模型多数以字为单位进行建模，无法感知粗粒度的领域词向量信息，对于实体嵌套词的识别效果并不好[7]。

针对上述问题，构建 LEBERT-CRF 实体识别模型，在 BERT 表示字向量的基础上利用词典适配器将课程领域词典的语义信息动态融入 BERT 编码器中，有效解决 BERT 模型以字向量对教育领域课程非结构化文本建模，不能感知粗粒度的领域词汇语义信息缺陷，优化实体识别的词向量表示方法。

1　相关工作

大量的深度神经网络技术应用于实体识别任务[8]，其中应用最广泛的就是 RNN[9]和长短期记忆网络(long-short term memory RNN，LSTM)。RNN 能够很好地学习序列样本的时序信息和相互关系，LSTM 的改进形式是在 RNN 中添加“门结构”以模拟人脑的记忆过程。在 LSTM 的基础上，其双向形式 BiLSTM[10]、基于注意力机制的 BiLSTM①陆续被提出。教育领域经典的命名实体识别方法是 BiLSTM-CRF[11]，即利用 BiLSTM 提取深度特征，用条件随机场(conditional random field，CRF)模型进行文本序列标注。随着预训练模型的提出，研究者将 BERT、XLNet 等预训练模型与 BiLSTM-CRF 结合，进一步提升

收稿日期：2023-09-16

基金项目：内蒙古自治区自然科学基金资助项目“利用软件演化历史识别与推荐重构克隆”(2018MS06009)；内蒙古自治区哲学社会科学研究专项资助项目“基于知识图谱的课程知识智能问答系统”(ZSZX21102)；内蒙古自治区自然科学基金联合资助项目“面向编程教育个性化学习的智能教育服务关键技术研究”(2023LHMS06009)。

作者简介：侯　敏(1973—)，女，副教授，主要从事软件工程，软件分析研究，E-mail:houmin920@163.com。

① ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 207-212.

命名实体识别的准确率[12]。Wei 等使用 BERT-BiLSTM-CRF 模型对教育领域突发事件语料库进行实体识别,准确率达到了 91.62%,为政府对教育突发事件的决策提供了帮助①。BERT-BiLSTM-CRF 模型结构见图 1。首先,通过 BERT 编码层获取文本对应的向量表示;其次,使用 BiLSTM 进行特征提取,获取文本中的上下文特征;最后使用 CRF 对序列进行解码和注释,以准确预测实体及其对应的类型。

上述基于神经网络的方法在应用到教育领域非结构化文本的实体识别时,需要先进行中文分词(Chinese word split,CWS)的处理,CWS 的准确程度对实体识别模型的性能至关重要[13]。依据不同的 CWS 方式将实体识别任务细分为基于字符粒度、基于词粒度的两种方式。通过查找大量文献发现,基于字符粒度的实体识别模型表现通常要比基于词粒度的效果好[14-15],原因是中文句子中词语的边界较为模糊,比较考验分词器的性能,基于词粒度的 CWS 切分难度较大,容易出现错误传播导致模型性能下降,且其表征字符的语义信息能力不强,因此研究者经常使用基于字符粒度的实体识别方法或者基于字符粒度的方法进行改进。

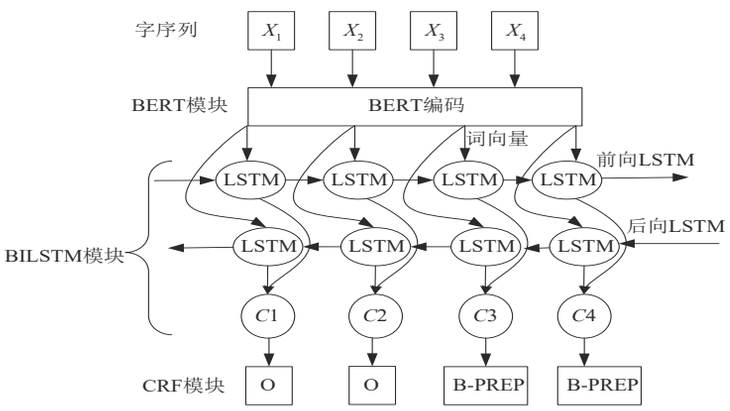

图 1 BERT-BiLSTM-CRF 实体识别模型

Fig. 1 BERT-BiLSTM-CRF entity recognition model

也有研究者专注于“字词融合”的方式优化词向量表示[16-17],该种方式在获取字粒度优秀特征表示的基础上融入了词粒度特征,并且经过多项实验证明要优于单纯以字进行切分的效果[18-19]。郭振东等[20]面向软件工程学科研究高质量的词向量生成方法,在基于字符粒度的向量表示基础上利用“首尾字求和”策略融合粗粒度的词向量表示,应用到分词任务中获得了较好的效果。但是该类方式关注的是文本中的全部特征,不能很好地捕捉句子中重点词汇的特征,对于嵌套语句的实体识别效果较差。因此,如何在字粒度向量的基础上捕获关键词向量的特征信息,是当下研究应重点考虑的问题。

与通用的自然语言文本不同,数据结构文本中包含大量的嵌套词,将其应用于实体识别模型存在边界难以界定的问题,例如“二叉树的先序遍历”一般会被识别为“二叉树”“先序遍历”,不能准确识别“二叉树的先序遍历”。而文本中字、词的语义表达对边界划分至关重要,因此本文基于字词融合的方式构建实体识别模型。

BERT 模型在实体识别任务中表现优异,因此本文选择基于 BERT 模型建模字向量的基础上实现字词融合方法。经过梳理文献,发现针对特定领域有两种基于 BERT 进行字词融合的方式,分别是模型级别的融合方式、BERT 底层特征融合的方式,见图 2。模型级别的融合方式是使用融合层(比如线性层)将 BERT 经过 12 层 Transformer 输出的字向量表示与词向量融合,接着输入推理层进行实体识别的方式,虽然该种方式能融合字符粒度和词粒度的语义表示信息,但该种方式更多是集中于浅层特征的融合,不能充分发挥 BERT 自身的序列建模优势,因此本文选择第二种方式,将数据结构课程的专业词汇融入 BERT 底层特征中,

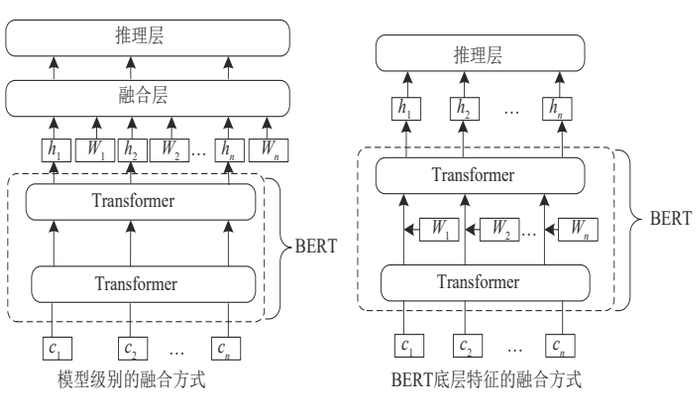

图 2 模型级别的融合方式、BERT 底层特征融合的方式

Fig. 2 Model-level and BERT bottom-level fusion method

① WEI K, WEN B. Named Entity Recognition Method for Educational Emergency Field Based on BERT[C]//2021 IEEE 12th International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2021: 145-149.

构建面向数据结构文本的实体识别模型。

以往的“字词融合”实体识别任务较为考验分词器的性能，对于文本中的嵌套词有多种分词歧义的情况，例如“二叉树的先序遍历”分词为“二叉树”“先序遍历”，该种情况忽略了“二叉树的先序遍历”作为整体的词义信息，而边界划分的好坏直接影响到实体识别的效果。因此，本文受文献[21]的启发提出一种字词融合的实体识别方法，融合字向量及其周围可能成词的向量表示，并根据词典适配器调整关键词的权重，优化原有字向量的语义表示，提升原有字向量的语义表达能力，降低边界划分错误对实体识别模型的影响。

2　LEBERT-CRF 实体识别模型

本文构建了 LEBERT-CRF 实体识别模型，将领域词向量融合到 BERT 的底层特征中，提升实体识别模型对嵌套词的识别效果。LEBERT-CRF 实体识别模型构建流程见图 3。首先，将课程文本句子 $S=\{c_1, c_2, \cdots, c_n\}$ 输入 BERT 模型，在 BERT 建模字向量的基础上，利用 word2vec 模型对匹配到的领域词汇 $\{ds_1, ds_2, \cdots, ds_n\}$ 建模，生成课程实体词级别的向量表示信息；其次，使用词典适配器将领域词汇的语义信息嵌入 BERT 编码器中，获得动态融合字符、词汇特征的词向量表示；最后，考虑标签的依赖性，使用 CRF 层进行解码，获得预测的标注序列，完成实体识别任务。LEBERT-CRF 的核心内容如下。

2.1　字符-领域词汇对序列

在命名实体任务中，BERT 模型多数以字向量为训练单位，导致词向量模型难以感知到专业词语整体的语义特征表示，无法生成高质量的领域文本词向量。针对教育领域课程文本的命名实体识别任务，将课程词典特征有效融入模型中，首先对字符系列进行扩充。给定一个课程词典 $\text{Dict}=\{d_1, d_2, \cdots, d_m\}$ 和包含 n 个字符的句子 $\text{Sen}_c=\{c_1, c_2, \cdots, c_n\}$，$d_i$ 为领域词典包含的词汇，c_i 为以字为单位进行切分的字符序列，通过匹配字符序列与课程词典进而获取句子中可能包含的领域词汇，记为 ds_i，以“结构体数组”举例，可能的分词情况有“结构体”“数组”以及“结构体数组”等多种，然后对于匹配到的词汇，将其分配于包含的字符，最后形成 cds 序列，$cds=\{(c_1, ds_1), (c_2, ds_2), \cdots, (c_n, ds_n)\}$。其中 c_i 表示句子中第 i 个字符，ds_i 表示分配给第 i 个字符的词汇，具体算法实现的伪代码如下。

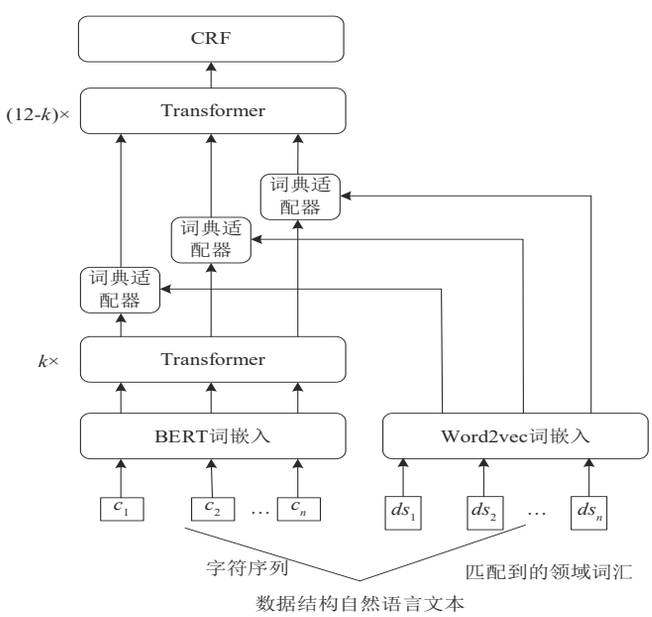

图 3　LEBERT-CRF 模型

Fig. 3　LEBERT-CRF entity recognition model

```
def sent_to_matched_words_boundaries(sent, lexicon_tree, max_word_num=None):
    """
    输入一个句子和词典树，返回句子中每个字所属的匹配词，以及该字的词边界
    字可能属于以下几种边界：
        B－：词的开始，0
        M－：词的中间，1
        E－：词的结尾，2
        S－：单字词，3
        BM－：既是某个词的开始，又是某个词中间，4
        BE－：既是某个词开始，又是某个词结尾，5
        ME－：既是某个词的中间，又是某个词结尾，6
```

```
        BME－：词的开始、词的中间和词的结尾，7

    Args：
        sent：输入的句子，一个字的数组
        lexicon_tree：词典树
        max_word_num：最多匹配的词的数量
    Args：
        sent_words：句子中每个字归属的词组
        sent_boundaries：句子中每个字所属的边界类型
    """
    sent_length = len(sent)
    sent_words = [[] for _ in range(sent_length)]
    sent_boundaries = [[] for _ in range(sent_length)]＃每个字符都有一个边界

    for idx in range(sent_length)：
        sub_sent = sent[idx:idx + lexicon_tree.max_depth]
        words = lexicon_tree.enumerateMatch(sub_sent)

        if len(words) == 0 and len(sent_boundaries[idx]) == 0：
            sent_boundaries[idx].append(3)＃S－
        else：
            if len(words) == 1 and len(words[0]) == 1：
                if len(sent_words[idx]) == 0：
                    sent_words[idx].extend(words)
                    sent_boundaries[idx].append(3)＃S－
            else：
                if max_word_num：
                    need_num = max_word_num － len(sent_words[idx])
                    words = words[:need_num]
                sent_words[idx].extend(words)
                for word in words：
                    if 0 not in sent_boundaries[idx]：
                        sent_boundaries[idx].append(0)＃S－
                start_pos = idx + 1
                end_pos = idx + len(word) － 1
                for tmp_j in range(start_pos, end_pos)：
                    if 1 not in sent_boundaries[tmp_j]：
                        sent_boundaries[tmp_j].append(1)＃M－
                    sent_words[tmp_j].append(word)
                if 2 not in sent_boundaries[end_pos]：
                    sent_boundaries[end_pos].append(2)＃E－
                sent_words[end_pos].append(word)
```

```
    assert len(sent_words) == len(sent_boundaries)

    new_sent_boundaries = []
    idx = 0
    for boundary in sent_boundaries:
        if len(boundary) == 0:
            print("Error")
            new_sent_boundaries.append(0)
        elif len(boundary) == 1:
            new_sent_boundaries.append(boundary[0])
        elif len(boundary) == 2:
            total_num = sum(boundary)
            new_sent_boundaries.append(3 + total_num)
        elif len(boundary) == 3:
            new_sent_boundaries.append(7)
        else:
            print(boundary)
            print("Error")
            new_sent_boundaries.append(8)
    assert len(sent_words) == len(new_sent_boundaries)
    return sent_words, new_sent_boundaries
```

2.2　词典适配器

如何将句子中每个位置的字符-词汇特征有效融入 BERT 模型中，是需要重点考虑的问题，本文采用词典适配器注入融合字词表示的特征。将字符-领域词汇对向量表示信息作为词典适配器的输入，形式化表示为 (h_i^c, l_i^{ds})，h_i^c 表示 BERT 中第 i 层 Transformer 输出的字向量，$l_i^{ds}=\{l_{i1}^d, l_{i2}^d, \cdots, l_{im}^d\}$ 表示与之对应的词向量，如公式(1)

$$l_{ij}^d = e^d(d_{ij}), \tag{1}$$

其中 e^d 是使用课程教材生成的训练语料，Word2vec 模型通过分句、jieba 结合领域词表分词进而生成词嵌入查找表，其中 d_{ij} 是 ds_i 中的第 j 个词汇。

通过对词向量进行非线性转换使之与字符向量对齐，如公式(2)

$$v_{ij}^d = W_2\left(\tanh\left(W_1 l_{ij}^d + b_1\right)\right) + b_2, \tag{2}$$

其中，W_1 和 W_2 分别是 $d_c \times d_w$ 和 $d_c \times d_c$ 的矩阵，b_1 和 b_2 表示偏置向量，d_c 代表 BERT 的隐藏层大小，d_w 表示词嵌入的维度。

考虑不同的词汇对任务的贡献度不同，采用如下方法处理：首先，将第 i 个字符相对应的 v_{ij}^d 表示为 $V_i=\{v_{i1}^d, v_{i2}^d, \cdots, v_{im}^d\}$，$V_i$ 维度为 (n, d_c)；其次，引入双线性注意力层计算词向量对应的注意力得分 a_i，同时对注意力权重和词向量进行加权求和得到 z_i^d；最后，将加权之后的词典特征 z_i^d 融入字符向量 h_i^c 中，得到 $\tilde{h}$，具体的计算方式为

$$a_i = \text{softmax}\left(k_i^c W_{\text{attn}} V_i^T\right), \tag{3}$$

$$z_i^d = \sum_{j=1}^{n} a_{ij} v_{ij}^d, \tag{4}$$

$$\tilde{h} = h_i^c + z_i^d。 \tag{5}$$

2.3 LEBERT 模型

LEBERT 由词典适配层和 BERT 构成,具体应用的过程见图 4。将词典适配器连接在 BERT 的 Transformer 层之间,将 Transformer 层输出的字向量与领域词汇的 embedding 相融合,进而获取高质量的词向量表示。

LEBERT 模型的处理过程如下:给定具有 n 个字符的句子 $S=\{c_1, c_2, \cdots, c_n\}$,对其扩展形成字符-领域词典的形式 $cds=\{(c_1, ds_1)(c_2, ds_2), \cdots, (c_n, ds_n)\}$,如 3.1 节所述。将 $\{c_1, c_2, \cdots, c_n\}$ 输入至嵌入层中,获取输出 $\{e_1, e_2, \cdots, e_n\}$,将其输入 Transformer 层中,计算方法为

$$G=LN(H^{t-1}+\mathrm{MHAttn}(H^{t-1})), \tag{6}$$

$$H^t=LN(G+\mathrm{FFN}(G)), \tag{7}$$

其中,H^t 表示 t 层的输出,LN 表示归一化,MHAttn 代表多头注意力机制,FFN 是两层前馈神经网络,其隐藏激活函数为 RELU。

为了在 k 层之后注入词汇信息,经过连续 k 层 Transformer 后输出字向量表示 $H^k=\{h_1^k, h_2^k, \cdots, h_n^k\}$,然后将 (h_i^k, l_i^{ds}) 通过词典适配器转换为 $\tilde{h}$,将 $\tilde{h}$ 继续输入至剩余的 $(12-k)$ 层 Transformer 中获得最终的输出,具体过程见图 4。

2.4 模型的训练与解码

模型解码层使用 CRF。首先,计算最后一层隐藏层输出 H^{12} 的得分,如公式(8)

$$O=W_oH^{12}+b_o。 \tag{8}$$

对于给定的序列标签 $y=\{y_1, y_2, \cdots, y_n\}$,对其概率进行定义,如公式(9)

$$p(y|s)=\frac{\exp(\sum_i(O_{i,y_i}+T_{yi-1,y_i}))}{\sum_{\tilde{y}}\exp(\sum_i(O_{i,\tilde{y}_i}+T_{\tilde{y}i-1,\tilde{y}_i}))}, \tag{9}$$

其中,T 代表转移得分矩阵,$\tilde{y}$ 代表所有的位置标签序列。

对于给定的 $\{s_j, y_j\}\|_{j=1}^N$,通过最小化句子级别的负对数似然损失训练模型,如公式(10)

$$L=-\sum_j\log(p(y|s)), \tag{10}$$

利用维特比算法找出分数最高的标签序列。

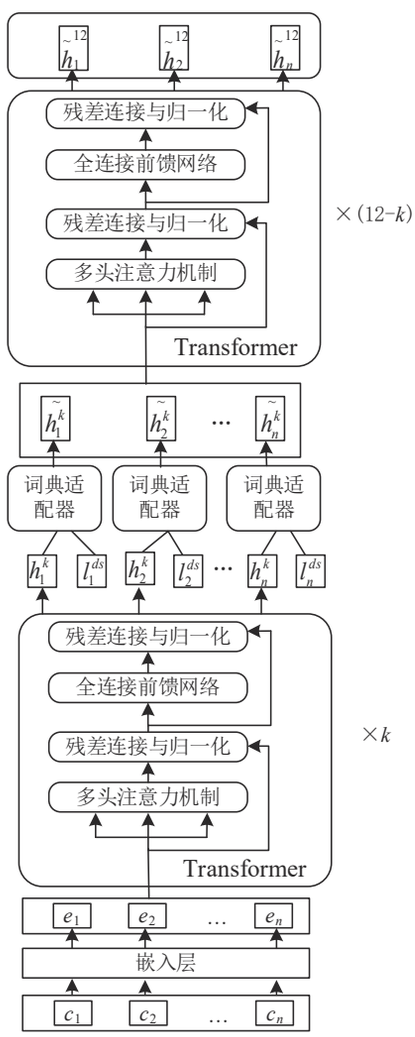

图 4 LEBERT 模型

Fig. 4 LEBERT Model

3 实验设计与分析

3.1 数据集

为了获取有价值的课程实体,验证实体识别模型的有效性,选取内蒙古师范大学计算机科学技术学院数据结构教学团队主编教材《数据结构(C 语言版)》作为数据集来源,该教材于 2021 年 6 月由科学出版社正式出版,已经在 2019－2021 级连续三届学生的教学中使用,共 20 万字左右。《数据结构(C 语言版)》课程文本中具有大量的领域专有名词,且存在较为普遍的嵌套命名实体情况,能够有效验证模型的有效性以及领域适用性。

在数据预处理阶段,需要对教材中的特殊符号、图片、表格、停用词以及 C 语言代码等内容进行删除,并使用“。”“?”“…”作为分句符号对数据集进行切分,总计 3 000 条句子,按 7∶2∶1 的比例切分训练集、测试集和验证集。通过领域专家人工标注的方式在教材中标注实体,共标注 424 个领域实体,其中“算法”类实体 94 个,“结

构”类实体 119 个，“名词术语”类实体 211 个，使用 BMES 的策略对需要标注的实体进行人工标注，“结构”类实例的词开头标注为 B-CON，中间为 M-CON，结尾为 E-CON，同理，“算法”类和“名词术语”类的实例分别用 ALG 和 PROP 替换，部分标注实例见表 1。

3.2　实验参数设置

实验中涉及的配置信息和参数见表 2－3。

3.3　评估指标

实体识别的主要评估标准是 P、R 和 $F1$，计算公式为

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

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

$$F1=\frac{2\times P\times R}{P+R}。 \tag{13}$$

P(precision)表示准确率，即所有被预测为正例的样本中，真正为正例的比例。R(recall)表示召回率，指的是在所有实际为正例的样本中有多少个被预测为正例。$F1$ 综合考虑 P 和 R 以评估模型的性能。TP、FP、FN 分别表示原本实例和预测实例均为正类、原本实例为负类，预测实例为正类、原本实例和预测实例均为负类的样本。

3.4　实验结果分析

3.4.1　*超参数实验*　超参数实验是在模型开始学习之前设置的，并非经过训练得到。为获取模型的最优性能，通常需要对超参数进行优化。考虑到不同参数值下模型表现各异，本研究设置超参数实验选取最优超参数。

在实体识别模型训练的过程中，学习率(leaning rate)和 batch_size 是影响模型性能的重要超参数。学习率代表每次权重参数更新的步长，其变化直接影响到模型收敛的状态。batch_size 表示每次传给模型训练样本的数量，其值的变化一般会对该模型泛化性能产生影响。

将标注后的数据转化为 json 格式输入至本研究构建的实验模型中，发现模型运行至第 5 轮的时候模型损失值会趋向稳定，模型逐渐收敛且验证集的精度最高，所以将轮次设定为 5 轮，以减少不必要的资源浪费。考虑到不同的 batch_size 和学习率对模型性能造成的影响，本文将 batch_size 分别设置为 4、8、16、32 和 64，学习率设置为 1e-5，2e-5，3e-5，4e-5，5e-5 和 1e-4，并且采取不同参数下的 $F1$ 进行对比见图 5。

由图 5 所示，当 batch_size 为 4、学习率为 2e-5 时模型的性能最优，$F1$ 为 94.71%。在模型的训练过程中，学习率是一个很重要的有监督参数，学习率过低会造成模型收敛速度慢、无法学习的问题，过高则会导致模型不能收敛，难以找到最优值的位置。通过观察发现在 batch_size 相同的情况下，在合理的范围内适当地增大学习率能够提升模型的收敛程度，帮助模型找到最优值。

在一定的范围内增加 batch_size 能够有效减少模型

表 1　实体识别部分标注实例

Tab. 1　Designation of entity types

实体类别	标注名称	示例
结构	CON	一维数组、图
算法	ALG	冒泡算法、快速排序算法
名词术语	PROP	主函数、指针域

表 2　配置信息

Tab. 2　Configuration information

名称	配置信息
操作系统	Ubuntu 18.04.5 LTS (GNU/Linux 5.4.0)
内存	45 GB
显存	20 GB
编程语言	Python
深度学习框架	pytorch1.6.0、tensorflow-gpu2.3.1
BERT 版本	bert-base-chinese

表 3　参数设置

Tab. 3　Parameter setting

超参数名称	内容
Transformer层数	12
hidden_size	768
head	12
词典适配器初始设置层数	1
输入序列最大长度	150

的训练时间,但提升 batch_size 意味着模型参数的更新次数减少,可能在结束训练之前模型还没有选择到最优的参数,会降低模型的泛化能力,进而影响模型的性能。由表 4 所示,当固定 leaning rate 为 2e-5 时,相比于其他的 batch_size,batch_size 为 4 模型整体表现是更优的,但其训练时间也最长。同时通过对比 batch_size=32、64 和 batch_size=4、8、16 的 $F1$ 值时,发现虽然较大的 batch size 整体 $F1$ 值不如取较小 batch_size 的 $F1$ 值,但在图 5 中,发现在固定 batch_size 的情况下随着学习率的增长其 $F1$ 值呈整体上升的趋势,因此在选择较大的 batch_size 时候 学习率也应相应地提高,以保证模型收敛的稳定性。

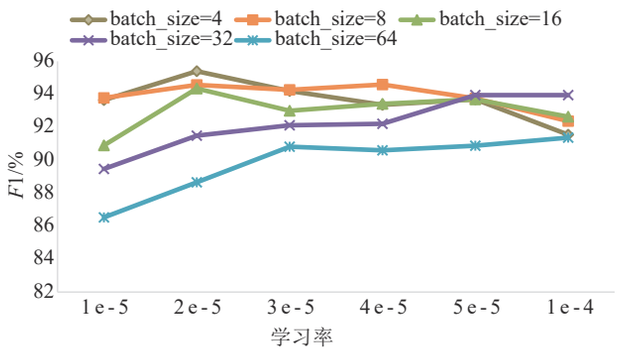

图 5 多个参数对应的 $F1$ 值

Fig. 5 $F1$ value corresponding to different hyperparameters

表 4 Batch_size 训练时间增量

Tab. 4 Training time increment table based on batch size

batch_size	总轮次	5轮的训练时间/s	时间增量/s	$F1$/%
4	5	338	0	94.71
8	5	224	−114	94.57
16	5	154	−70	94.35
32	5	124	−30	91.49
64	5	108	−16	88.66

3.4.2 LEBERT 模型优化 LEBERT 是在 BERT 对字序列建模的基础上置入了词典适配器,考虑到 BERT 的不同层对句法特征和语义特征的学习效果不同[22],在置入词典适配器学习字词融合的特征偏重也不同,会对模型的下游任务造成一定的影响。为探究 BERT 的 Transformer 不同层之后置入词典适配器的实验效果,设定 batch_size 为 4,learning rate 为 2e-5。

选择在 1,3,6,9,12 层之后、多层或者所有层之后置入词典适配器,见表 5。仅从单层的性能而言,词典适配器在第 9 层的 Transformer 之后取得最好的模型性能,表明 BERT 中间层的词典特征融合促进了 BERT 的字符特征与领域词汇特征之间的交互,接着优先选取准确率高的层进行融合,设置 9 层、12 层融合,1 层、9 层、12 层融合,1 层、3 层、9 层、12 层融合以及所有层的融合,发现第 9 层和第 12 层融合的效果最好,验证了 LEBERT 设置中间层或者更深层次之间进行特征融合会有更好的表现,而所有层融合的效果不是很好,原因是融合多层特征会导致模型过拟合。

表 5 不同层特征融合对比的实验结果

Tab. 5 Experimental results of lexicon adapters inserted after different transformer layers

不同层特征融合	层数	$F1$/%
单层特征	1	94.71
	3	94.40
	6	94.20
	9	94.91
	12	94.73
多层特征	9,12	95.47
	1,9,12	94.96
	1,3,9,12	94.74
所有层特征	1−12	88.39

3.4.3 几种实体识别模型的对比实验结果分析 为验证字词融合方式的 LEBERT-CRF 实体识别模型性能。首先,通过对比其他经典的以字为切分单位的主流命名实体识别模型 BILSTM-CRF、BERT-CRF、BERT-BILSTM-CRF 验证本文选择“字词融合”实体识别方式的有效性;其次,选择目前“字词融合”方式表现较好的 BERT+Word 实体识别模型作为对比模型进行实验结果分析,验证本文模型较主流的字词融合模型表现更好。

融合字词表示模型普遍要比以字向量进行训练的实体识别模型 BILSTM-CRF,BERT-CRF 以及 BERT-BILSTM-CRF 准确率高(表 6),证明字词融合的语义表示效果优于单纯以字向量进行切分的效果,同时采用 BERT 自身深层特征融合与模型级字词特征融合效果,设置当前实体识别任务中字词融合准确率较高的模型融合方式 BERT+Word 作为对比模型,与本文采用的实验模型相比,BERT+Word 将 BERT12 层的 Transfomer 向量特征表示与句子输入直接转换的词典向量拼接并输出至 BILSRM-CRF 层进行推理的方式,可以明显看出本文提出的模型效果优于 BERT+Word 模型。

通过对比 BERT＋Word 和 LEBERT-CRF 的一些识别实例进行分析，发现本文提出的模型对于嵌套实体词的识别效果较好（表 7），例如表 7 实例 1 中的“集合结构”一词，BERT＋Word 模型只能识别“集合”一词，而本文提出的模型则能有效识别“集合结构”，表 8 实例 2 同理证明 LEBERT 的中间层和较高层的字词特征融合对于词汇边界的划分效果更好。

表 6　不同模型的实验结果对比

Tab. 6　Analysis of experimental results for different models

模型名称	词向量	*P*	*R*	*F*1
BILSTM-CRF	字向量	90.06	90.17	90.11
BERT-CRF	字向量	92.68	92.94	92.81
BERT-BILSTM-CRF	字向量	94.65	94.26	94.45
BERT＋Word	字词融合	93.97	95.23	94.60
LEBERT-CRF	字词融合	94.72	96.23	95.47

表 7　实例分析 1

Tab. 7　Example analysis 1

实体词（段落文本）	匹配的词	字符序列	标签	BERT＋Word 预测结果	LEBERT-CRF 预测结果
通常有四类基本逻辑结构：集合结构……	逻辑结构，集合，结构，集合结构	通常有四类基本逻辑结构：集合结构	O O O O O O O B-CON I-CON I-CON E-CON O B-CON I-CON I-CON E-CON	O O O O O O O B-CON I-CON I-CON E-CON O B-CON E-CON O O	O O O O O O O B-CON I-CON I-CON E-CON O B-CON I-CON I-CON E-CON

表 8　实例分析 2

Tab. 2　Example analysis 2

实体词（段落文本）	匹配的词	字符序列	标签	BERT＋Word 预测结果	LEBERT-CRF 预测结果
改为循环结构的阶乘问题算法	循环，结构，循环结构，算法	改为循环结构的阶乘问题算法	O O B-CON I-CON I-CON E-CON O O O O O O B-PROP E-PROP	O O B-CON E-CON O O O O O O O B-PROP E-PROP	O O B-CON I-CON I-CON E-CON O O O O O O B-PROP E-PROP

4　结论

针对教育领域课程文本中出现的实体嵌套问题，提出一种基于 LEBERT-CRF 的实体识别模型。在原有基于 BERT 编码字向量的基础上融合其附近可能成词的向量表示，并根据词典适配器调整关键词的权重，优化原有字向量的语义表示，实验结果表明该模型有效降低分词歧义导致的错误传播问题，能够很好地适应课程文本的实体识别任务。

参考文献：

[1]　王颖洁，张程烨，白凤波，等．中文命名实体识别研究综述[J]．计算机科学与探索，2023，17(2)：324-341.

[2]　赵宇博，张丽萍，闫盛，等．个性化学习中学科知识图谱构建与应用综述[J]．计算机工程与应用，2023，59(10)：1-21.

[3]　LAMPLE G，BALLESTEROS M，SUBRAMANIAN S，et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv，2016：1603.01360.

[4]　杜鹏，张有明，朱郑州，等．迁移学习在低资源场景实体识别中的应用研究[J]．计算机科学与探索，2023，17(4)：912-921.

[5]　DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv，2018：1810.04805.

[6]　谢腾，杨俊安，刘辉．基于 BERT-BiLSTM-CRF 模型的中文实体识别[J]．计算机系统应用，2020，29(7)：48-55.

[7]　赵丹丹，黄德根，孟佳娜，等．多头注意力与字词融合的中文命名实体识别[J]．计算机工程与应用，2022，58(7)：142-149.

[8]　张汝佳，代璐，王邦，等．基于深度学习的中文命名实体识别最新研究进展综述[J]．中文信息学报，2022，36(6)：20-35.

[9] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011(12): 2493-2537.
[10] WÖLLMER M, EYBEN F, GRAVES A, et al. Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework[J]. Cognitive Computation, 2010, 2(3): 180-190.
[11] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv, 2015: 1508.01991.
[12] YAN R, JIANG X, DANG D. Named entity recognition by using XLNet-BiLSTM-CRF [J]. Neural Processing Letters, 2021, 53(5): 3339-3356.
[13] 殷章志,李欣子,黄德根,等 . 融合字词模型的中文命名实体识别研究[J]. 中文信息学报,2019,33(11):95-100.
[14] 刘振元,许明阳,王承涛 . 基于数据增强和字词融合特征的实体槽位识别[J]. 华中科技大学学报(自然科学版),2022,50(11):101-106.
[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.
[16] 李书琴,张明美,刘斌 . 融合字词语义信息的猕猴桃种植领域命名实体识别研究[J]. 农业机械学报,2022,53(12):323-331.
[17] 邓依依,邬昌兴,魏永丰,等 . 基于深度学习的命名实体识别综述[J]. 中文信息学报,2021,35(9):30-45.
[18] 尹成龙,陈爱国 . 融合多重嵌入的中文命名实体识别[J]. 中文信息学报,2023,37(4):63-71.
[19] 杨振平,毛存礼,雷雄丽,等 . 融入词集合信息的跨境民族文化实体识别方法[J]. 中文信息学报,2022,36(10):88-96.
[20] 郭振东,林民,李成城,等 . 基于 BERT-CRF 的领域词向量生成研究[J]. 计算机工程与应用,2022,58(21):156-162.
[21] LIU W, FU X, ZHANG Y, et al. Lexicon enhanced chinese sequence labeling using bert adapter[J]. arXiv preprint arXiv,2021:2105.07148.
[22] ROGERS A, KOVALEVA O, RUMSHISKY A. A primer in BERTology: What we know about how BERT works [J]. Transactions of the Association for Computational Linguistics, 2021(8): 842-866.

Research on Named Entity Recognition of Data Structure Text Based on Domain Word Vector Fusion

HOU Min, GAO Mao, ZHANG Liping, YAN Sheng, ZHAO Yubo

(*College of Computer Science and Technology*, *Inner Mongolia Normal University*, *Hohhot* 010022, *China*)

Abstract: The BERT pre training model, which uses words as segmentation units, performs well in entity recognition tasks, but it ignores coarse-grained domain vocabulary as the overall semantic representation information, is not effective for recognizing a large number of nested entities in educational curriculum texts. To address the above issues, a dynamic fusion word and word level word vectors for LEBERT-CRF education domain course text entity recognition method is proposed. In the method, the dictionary adapter is used to efficiently integrate domain dictionary features into the BERT model and thereby improving the recognition effect of the BERT model on entity word boundaries and better adapting to course knowledge extraction tasks. The experimental results indicate that the LEBERT-CRF model performs better than other mainstream entity recognition models do, with an F1 of 95.47%.

Key words: entity recognition; LEBERT; domain word vector; word fusion

【责任编辑 张颖娟】