计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000154

基于多头注意力机制与词典特征融合的招标文件命名实体识别算法

杨 华 王宝会
北京航空航天大学软件学院 北京 100191
(987754124@qq.com)

摘 要 招标文件的编制和审核,是确保招标过程顺利进行的重要环节。实体识别技术在招标文件审核过程中可以显著提高信息提取的准确性和效率,增强信息的可读性和可检索性。但招标文件内容复杂,专业术语多,长实体识别难度大,传统命名实体识别方法在此类任务中的表现欠佳。为此,提出了一种命名实体识别技术,该技术整合了多头注意力机制、词汇特征融合以及基于RoBERTa的BiLSTM-CRF模型,简称为RoBERTa-DFF-BiLSTM-MHA-CRF。此方法利用RoBERTa模型作为基础输入层,有效提升了对长距离依赖特征的识别能力;通过引入多头自注意力机制,进一步增强了对长跨度实体的识别能力;融合领域专业术语的词典特征,解决了专业术语边界不明显的问题。实验结果表明,该模型在招标文件的命名实体识别任务中显著提升了信息提取的准确性和效率,相较于BERT-BiLSTM-CRF,在Precision上提升了2.49个百分点,在Recall上提升了4.28个百分点,在F1上提升了3.37个百分点,降低了时间和人力成本,为招投标文件的信息提取提供了一种高效的新方案。

关键词:招标文件实体识别;多头注意力机制;词典特征融合;Roberta
中图分类号 TP301

Bidding Document Named Entity Recognition Algorithm Based on Multi-head Attention Mechanism and Dictionary Feature Fusion

YANG Hua and WANG Baohui
School of Software,Beihang University,Beijing 100191,China

Abstract The preparation and review of bidding documents play a crucial role in ensuring the smooth operation of the bidding process. Entity recognition technology can notably enhance the accuracy and efficiency of information extraction,thereby improving the readability and retrievability of information during the review of bidding documents. However,due to the complexity of the content and the presence of numerous specialized terms,recognizing long entities poses a significant challenge. Traditional methods for named entity recognition(NER) perform poorly in addressing these issues. This paper proposes an NER approach named Roberta-DFF-BiLSTM-MHA-CRF,which integrates a multi-head attention mechanism,dictionary feature fusion,and the Roberta-BiLSTM-CRF model. Utilizing Roberta as the input layer,this method enhances the capability to capture long-range dependencies. The introduction of the multi-head self-attention mechanism improves the recognition of long entities. Meanwhile,incorporating domain-specific dictionary features addresses the issue of unclear term boundaries. Experimental results demonstrate that the proposed model significantly boosts the accuracy and efficiency of information extraction in the context of NER for bidding documents. When compared to the Bert-BiLSTM-CRF model,it achieves a 2.49 percentage point improvement in precision,a 4.28 percentage point increase in recall,and a 3.37 percentage point enhancement in F1 score. These improvements effectively reduce time and labor costs,offering an efficient new solution for information extraction from bidding documents.

Keywords Entity recognition in tender documents,Multi-head attention,Dictionary feature fusion,Roberta

1 引言

招标文件的编制和审核,是确保招标过程顺利进行的重要环节。通过明确招标需求,编写详细的招标文件,进行多维度的内部和外部审核,可以确保招标文件的完整性和合规性[1]。实体识别技术在招标文件审核过程中可以显著提高信息提取的准确性和效率,增强信息的可读性和可检索性。

实体识别技术可以自动提取招标公告中的关键实体,如项目名称、招标方、投标截止日期、资格要求等,减少人工操作,提高信息提取的准确性和效率。另外,自动化处理可以减少因人工操作不当导致的错误,确保信息的完整性和准确性。

命名实体识别(Named Entity Recognition,NER)最初采用了基于规则和字典的方法,这种方法对于具有一定模式的句子能够提供较为准确的识别,但其高度依赖人工设定的规则,不仅成本高昂,而且泛化能力较差。随后,一系列基于统计机器学习的技术涌现,如最大熵模型[2]、条件随机场(Conditional Random Field,CRF)[3]、隐马尔可夫模型(Hidden Markov Model,HMM)[4]等。这些方法尽管减轻了对人工规则的依赖,但仍然需要通过大规模的标注语料库来确保模型的有效训练。随着深度学习技术的不断进步,特别是卷积神

通信作者:王宝会(wangbh@buaa.edu.cn)

经网络(Convolutional Neural Networks, CNN)[5]、循环神经网络(Recurrent Neural Network, RNN)[6]、长短期记忆网络(Long Short-Term Memory, LSTM)[7]、双向长短期记忆网络(Bi-directional Long-Short Term Memory, BiLSTM)[8]等迅猛发展,命名实体识别研究取得了显著进展。这些基于深度学习的方法不仅突破了传统方法的局限性,还有效解决了数据维度高、信息冗余等问题,能够在较少的人工干预下实现更优的识别性能。

然而,招标文件中内容比较复杂,专业术语较多,含有较多的长实体。因此,本文提出了一种结合多头注意力机制、词典特征融合与 Roberta-BiLSTM-CRF 的面向招标领域的命名实体识别方法(Roberta-DFF-BiLSTM-MHA-CRF),该方法不仅能够有效减少时间和人力成本,而且能够解决长实体识别和专业术语边界的问题,为招投标文件信息提取问题提供了一种高效的新方案。

本文的主要创新点如下:

1)针对 BERT 对于招标文件准确率不高的问题,采用 RoBERTa 代替 BERT 作为输入层,能够更好地捕捉长距离依赖关系,提高实体识别的准确性;

2)针对招标文件中较长实体识别效果不好的问题,引入多头自注意力机制以捕捉语义相关信息的多重特征,有助于识别跨句的实体;

3)针对中文词语边界不明显的问题,引入领域专业术语的词典特征融合,提高模型的性能。

2 相关研究工作

命名实体识别对语义分析和应用至关重要。目前,在许多开放的数据集上,实体识别技术已经展现了卓越的性能。因此,越来越多的研究人员尝试将此技术应用于专业领域,如生物医学领域[9]、军事领域[10]、电力领域[11]等。Luo 等[12]采用基于 RoBERTa 预训练模型、跨度识别和对抗训练的标签指针网络融合深度模型,解决了中文军事领域文本中的命名实体识别问题。Li 等[13]使用基于全词掩码的 BERT 预训练模型和 BiLSTM 结合注意力机制的方法,解决了电力设备缺陷文本中的实体识别问题,提高了模型在地点、缺陷内容和设备 3 种实体上的识别性能。

但在招标领域,命名实体识别技术的应用目前相对较少。Zhang 等[14]构建了一种先进的命名实体识别模型,该模型结合了 BiLSTM 和 CRF,解决了提取招标人、招标代理及招标编号 3 类实体信息的问题。Mi 等[15]通过结合 CNN 提取的汉字五笔编码和 BERT 获取的字符特征,解决了招标领域中物料数据书写不规范、分词错误、词语多义性以及未充分考虑中文特有的字形特征等问题。Aejas 等[16]通过创建高质量的合同命名实体识别数据集并评估多种深度学习模型,发现 Contracts-BERT-base 模型在实体识别任务中表现最佳。Ahmet 等[17]提出了一种结合 BERT-base-cased 分类模型和规则方法的命名实体识别算法,专门用于金融文档的自动验证。Ma 等[18]使用融合了 BERT 预训练的 BiLSTM-CRF 深度学习模型,解决了物流标书中招标公司、运输地点和运输项目 3 类实体信息的自动提取问题。最近的研究主要集中在结合 BERT, BiLSTM 和 CRF 的模型上,这种组合在多个领域的命名实体识别任务中表现出色。Zhang 等[14]和 Ma 等[15]的研究分别解决了商情文本和物料数据中的命名实体识别问题,显著提升了实体识别的性能。此外,最近的研究工作表明,通过微调这些大型模型或采用提示工程(Prompt Engineering)的方法,可以将实体识别问题转化为问答(QA)或填空问题,从而在多个领域展现卓越的性能。这种方法尽管在某些情况下表现出色,但在特定领域中可能面临挑战,例如对专业术语的识别和长实体的边界划分。本研究旨在通过引入多头注意力机制和词典特征融合,进一步提升基于 RoBERTa 的 BiLSTM-CRF 模型在招标文件命名实体识别任务中的性能,并进行对比分析。

3 Roberta-DFF-BiLSTM-MHA-CRF 模型

Roberta-DFF-BiLSTM-MHA-CRF 模型主要包含 5 个部分:Roberta 词向量编码层、词典特征融合层、BiLSTM 层、多头注意力机制及 CRF 层。首先,将已标注的文本序列输入 Roberta 层获得相应的字向量,在 Roberta 的输出之后插入词典特征,将词典特征与 Roberta 模型的输出特征进行融合,再将融合后的字向量输入 BiLSTM 层,对序列特征进行学习。然而,利用多头注意力机制对 BiLSTM 提取的隐含特征进行权重调节。最后,由 CRF 层给出综合得分并预测结果。模型结构如图 1 所示。

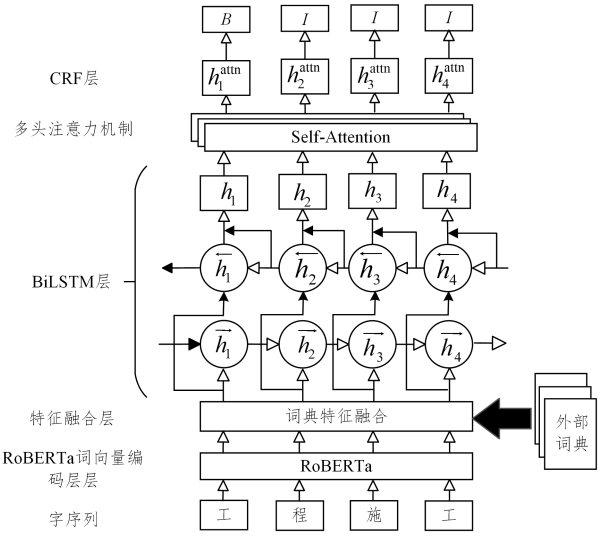

图 1 命名实体识别模型网络结构图

Fig. 1 Network structure diagram of named entity recognition model

3.1 输入表示层 Roberta

Roberta 是 BERT 模型的一种改进版本,由 Facebook AI 在 2019 年提出[19]。Roberta 在 BERT 的基础上进行了优化,在训练方法上移除下句预测任务并引入动态掩码,在模型参数量、batch size 和训练数据上规模更大。其模型结构如图 2 所示。

在 Roberta 模型中,Input 为原始文本的输入,在文本的开头和句子之间分别插入起始标记[CLS]和分隔标记[SEP],采用字向量、句向量和位置向量的 Embedding 编码之和作为特征向量。假设输入文本以特殊符号"[CLS]"开始,该文本被表示为一系列元素的序列$[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$,$n$ 代表模

型预设的输入句子的最大长度。那么，Roberta 模型会将每个x_i编码为向量$\boldsymbol{E}_i$：

$$\boldsymbol{E}_i=\boldsymbol{E}_{\text{token}}(x_i)+\boldsymbol{E}_{\text{seg}}(x_i)+\boldsymbol{E}_{\text{pos}}(x_i) \tag{1}$$

其中，$\boldsymbol{E}_{\text{token}}(x_i)$，$\boldsymbol{E}_{\text{seg}}(x_i)$及$\boldsymbol{E}_{\text{pos}}(x_i)$分别表示字向量、句向量和位置向量。此外，随机选择 15%的词汇作为掩码词，每次将序列输入模型时，这些掩码词汇都会重新选定并应用以下遮挡策略进行替换。1)80%的情况下用[MASK]替换；2)10%的情况下保持不变；3)另外 10%随机替换为其他词汇。接着，将$[\boldsymbol{E}_1,\boldsymbol{E}_2,\cdots,\boldsymbol{E}_n]$送入双向 Transformer 网络中，并加载预训练好的模型参数。经过网络的处理，得到输入序列的最终向量表示$[\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n]$。

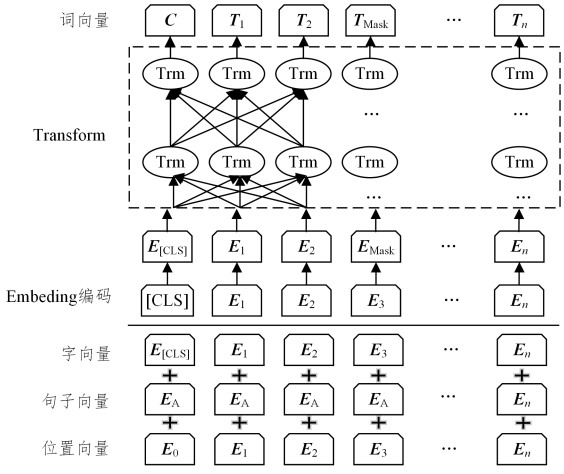

图 2　Roberta 模型结构

Fig. 2　Roberta model structure

3.2　词典特征融合层

为了减少分词错误对下游任务的负面影响，提高中文文本中的实体识别效果，本文采用特征融合方法，利用专业术语词典为模型提供精确的词语语义和边界特征。词典特征是指从预定义的词典中提取的特征，用于增强模型的表示能力。在命名实体识别(NER)任务中，词典特征可以包括二进制特征、词频特征和词语的位置信息等。

假设有一个词典文件$D=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}$，其中，$D$ 是词典集合，$\boldsymbol{w}_i$是词典中的第 i 个词，该词典包含了一些特定的实体词汇。对于给定的 Token $\boldsymbol{t}_i$，其对应的二进制特征 $f(\boldsymbol{t}_i)$可以定义为：

$$f(\boldsymbol{t}_i)=\begin{cases}1, & \text{if } t_i\in T\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

其中，$\boldsymbol{t}_i$表示句子中的第 i 个 Token，T 是包含所有领域特定术语的集合(即词典)。如果 t_i属于术语集合 T，则 $f(\boldsymbol{t}_i)=1$；否则，$f(\boldsymbol{t}_i)=0$。

词频特征指每个词在文本中出现的次数。可以通过式(3)统计词频：

$$f(\boldsymbol{t}_j)=\sum_{i=1}^{m}\mathbb{I}(\boldsymbol{t}_i=\boldsymbol{t}_j) \tag{3}$$

其中，$\mathbb{I}$是指示函数，当条件$\boldsymbol{t}_i=\boldsymbol{t}_j$成立时返回 1，否则返回 0。对于给定的文本 $T=\{\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_m\}$，词频特征$\boldsymbol{f}_j$可以定义为：

$$\boldsymbol{f}_j=\boldsymbol{f}(\boldsymbol{t}_j) \tag{4}$$

将词典特征与 BERT 模型的输出特征进行融合，通常使用拼接(Concatenation)的方式。假设 $\boldsymbol{S}\in\mathbb{R}^{b\times m\times d}$是 BERT 模型的输出特征，其中 b 是批次大小，m 是序列长度，d 是特征维度；二进制特征表示为 $\boldsymbol{B}\in\mathbb{R}^{b\times m}$，词频特征表示为 $\boldsymbol{F}\in\mathbb{R}^{b\times m}$。为了将这些特征与 BERT 的输出特征 S 拼接起来，需要先将它们扩展到与 S 相同的维度。扩展特征的维度的方式可由式(5)表示。

$$\begin{cases}\boldsymbol{B}'=\boldsymbol{B}\otimes\boldsymbol{1}_d\\ \boldsymbol{F}'=\boldsymbol{F}\otimes\boldsymbol{1}_d\end{cases} \tag{5}$$

其中，$\otimes$表示外积操作，$\boldsymbol{1}_d$是一个 d 维的全 1 向量，使得 $\boldsymbol{B}'$和 $\boldsymbol{F}'$的形状变为$\mathbb{R}^{b\times m\times d}$。拼接特征的方式如下：

$$\boldsymbol{S}'=\text{concat}(\boldsymbol{S},\boldsymbol{B}',\boldsymbol{F}') \tag{6}$$

其中，concat 代表将向量在最后一个维度上拼接的操作，使得 $\boldsymbol{S}'$的形状变为$\mathbb{R}^{b\times m\times(d+2)}$。

3.3　特征提取层 BiLSTM

LSTM 是 RNN 的一种改进方式[20]。LSTM 的神经元中引入了门函数，通过 3 个门结构(遗忘门重置数据、输入门读取数据、输出门输出数据)实现对前后文信息的取舍。LSTM 的结构如图 3 所示。

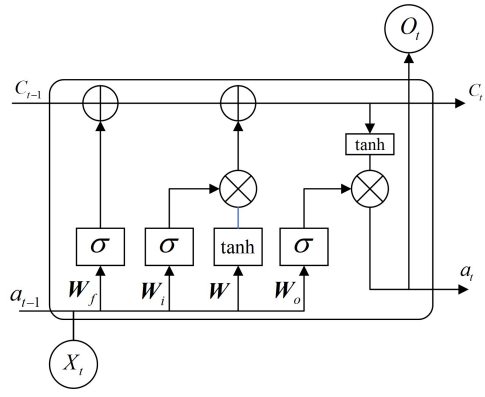

图 3　LSTM 的神经元

Fig. 3　LSTM neuron

每个神经元都有 3 个输入与 3 个输出，$\boldsymbol{X}_t$ 是该时刻新加入的信息，$\boldsymbol{a}_{t-1}$与$\boldsymbol{c}_{t-1}$是上文信息的表示。首先，将 $\boldsymbol{X}_t$与 $\boldsymbol{a}_{t-1}$合并后复制为 4 份，其中 3 份分别与遗忘门、输入门和输出门对应的权重矩阵$\boldsymbol{W}_f$，$\boldsymbol{W}_i$，$\boldsymbol{W}_o$进行矩阵运算。随后，通过激活函数的非线性变换，得到遗忘门权重、输入门权重和输出门权重；最后一份放入 RNN 中计算。由于一个 LSTM 只能处理单向，为了更加准确地分析语句，将两个输入方向相反的 LSTM 模型组合为 BiLSTM 使用，如图 4 所示。

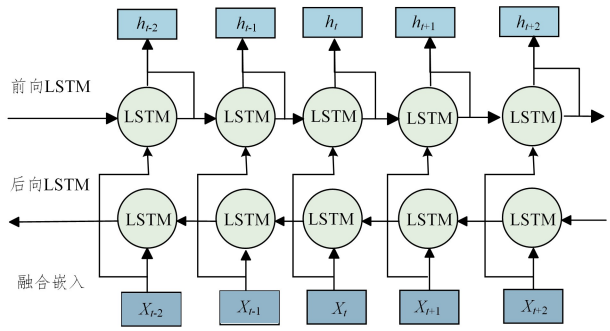

图 4　BiLSTM 的结构图

Fig. 4　BiLSTM structure diagram

3.4　多头注意力机制

预训练语言模型 Roberta 能够学习通用的语言特征并适

用于多种任务，但是可能不足以满足特定领域（如招标文件命名实体识别）的需求。为此，引入额外的注意力层不仅可强化对领域特有术语和长实体的捕捉，还能在更深层次上理解和建模复杂的依赖关系，特别是对于那些涉及长距离信息关联的复杂句式[21]。

多头自注意力机制通过并行运行多个注意力“头”来实现，每个头独立计算缩放点积注意力，然后将这些头的输出合并，并通过线性层转换，形成最终的表示维度[22]。这一机制的计算步骤如下。

第一步 对 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别进行线性映射：

$$\boldsymbol{Q}'=\boldsymbol{Q}*\boldsymbol{W}_j^Q \tag{7}$$

$$\boldsymbol{K}'=\boldsymbol{K}*\boldsymbol{W}_j^K \tag{8}$$

$$\boldsymbol{V}'=\boldsymbol{V}*\boldsymbol{W}_j^V \tag{9}$$

在多头自注意力模型中，每个头都通过训练获得自己的权重矩阵 $\boldsymbol{W}_j^Q$，$\boldsymbol{W}_j^K$ 和 $\boldsymbol{W}_j^V$，并且这些矩阵通常具有相同的维度 H。

第二步 计算缩放点积注意力。

$$\boldsymbol{M}_j=\mathrm{softmax}\left(\frac{\boldsymbol{Q}'\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}' \tag{10}$$

其中，$\boldsymbol{Q}'$ 是经过线性变换的查询矩阵；$\boldsymbol{K}'$ 是经过线性变换的键矩阵；$\boldsymbol{V}'$ 是经过线性变换的值矩阵；$1/\sqrt{d_k}$ 是一个缩放因子，用来缩小点积结果的规模，防止在应用 softmax 函数时梯度消失；d_k 是查询和键的维度。最后，使用注意力权重矩阵对值矩阵$\boldsymbol{V}'$进行加权求和，得到输出矩阵$\boldsymbol{M}_j$。这个输出矩阵包含了根据注意力权重加权后的值向量，反映了输入序列中不同部分对当前查询的重要性。

3.5 标签解码层 CRF

CRF 是一种统计学上的序列标注方法[23]。它将序列标注任务看作一个多类别分类问题，其中类别数为标签数量的 n 次方，n 为序列长度。CRF 不仅捕捉了序列中的上下文信息，还建模了标签之间的转移概率，这使得模型能够从整体上考虑序列，寻找最优的标签序列。CRF 计算条件概率的方式、得分函数、标签概率如下：

$$P(\boldsymbol{y}_1\cdots\boldsymbol{y}_n\mid\boldsymbol{x}_1\cdots\boldsymbol{x}_n)=P(\boldsymbol{y}_1\cdots\boldsymbol{y}_n\mid\boldsymbol{x}) \tag{11}$$

$$\boldsymbol{x}=(\boldsymbol{x}_1\cdots\boldsymbol{x}_n) \tag{12}$$

$$P(\boldsymbol{y}\mid\boldsymbol{X})=\frac{\mathrm{e}^{s(\boldsymbol{X},\boldsymbol{y})}}{\sum\limits_{\tilde{\boldsymbol{y}}\in Y_{\boldsymbol{X}}}\mathrm{e}^{s(\boldsymbol{X},\tilde{\boldsymbol{y}})}} \tag{13}$$

其中，$\boldsymbol{y}_i$ 代表第 i 个字符的标签，$\boldsymbol{x}_i$ 代表第 i 个字符；$\mathrm{s}(\boldsymbol{X},\boldsymbol{y})$ 为得分函数（用于衡量输入序列 $\boldsymbol{X}$ 与标签序列 $\boldsymbol{y}$ 的匹配度）；分母为归一化项（配分函数），表示对所有可能的标签序列 $\tilde{\boldsymbol{y}}$ 的得分函数进行指数化求和；$\boldsymbol{Y}_{\boldsymbol{X}}$ 为输入序列 $\boldsymbol{X}$ 对应的所有可能标签序列的集合。

4 实验与分析

4.1 数据集

实验所采用的数据集源自某招标集团，用于对招标文件进行分类标注。该数据集涵盖了九大类招标文件，共包含 269 篇文档。文档中标注了 80 种不同的实体，主要包括项目名称、项目编号、候选人名称、计划工期、公示时间、采购人名称、采购代理机构等多种实体类型。数据集类型及数量如表 1 所列。

表 1 数据集的类别及数量

Table 1 Categories and quantities of the dataset

类型	数量
不受两法约束的服务	19
不受两法约束的工程	22
不受两法约束的货物	31
基于招标投标法的工程服务	64
基于招标投标法的工程货物	22
基于招标投标法的工程施工	55
基于政府采购法的服务	16
基于政府采购法的工程	22
基于政府采购法的货物	18

将以上数据按照预设长度进行划分，划分后的数据集共 1 194 条。为了进行模型的训练验证，将这 1 194 条数据按照 8∶1∶1 的比例分为训练集、验证集和测试集 。其中，训练集包含 956 项数据，用于模型的学习和参数优化；验证集与测试集各包含 119 项数据，验证集旨在评估模型性能，测试集用于测试模型性能。

4.2 参数设置

通过对验证集进行多次实验对比和细致的参数调整，评估不同超参数设置下模型的性能，并选择了在验证集上表现最佳的参数配置。具体参数设置如表 2 所列。

表 2 实验参数设置

Table 2 Experimental parameter settings

参数	值
学习率	3×10^{-5}
Batch 大小	4
Epoch 数量	75
LSTM 单元	128
优化函数	Adam
Dropout 率	0.1
多头注意力机制头数	16

4.3 评价指标

NER 是一项核心任务，涉及在文本中定位实体并分类识别其类型。为了准确衡量模型在 NER 任务上的表现，本研究采用了 3 个核心评价指标：精确率（Precision）、召回率（Recall）和 F1 分数（F1-score）。精确率用于评估模型预测的准确性，通过式(14)计算得到：

$$P=\frac{TP}{TP+FP} \tag{14}$$

其中，TP（True Positives）表示模型成功识别为正例的样本数，而 FP（False Positives）表示模型错误地将负例预测为正例的样本数。

召回率用于衡量模型的全面性，通过式(15)计算：

$$R=\frac{TP}{TP+FN} \tag{15}$$

其中，FN（False Negatives）代表未被识别出的实体数量。

F1 是精确率和召回率的调和平均值，其计算式如式(16)所示：

$$F1=\frac{2\times P\times R}{P+R} \tag{16}$$

F1 分数平衡了精确率和召回率，是评估模型整体性能的有效指标。这些指标共同为模型在 NER 任务上的表现提供了全面的评估，确保模型在识别实体时既准确又全面。通过这些量化指标，可以对不同模型的性能进行比较和优化。

4.4 模型对比实验

本研究对比了多种深度学习模型在招标领域命名实体识

别任务上的表现。实验涉及的模型包括 BERT，BERT-BiGRU-CRF，BERT-BiLSTM-CRF，Roberta-BiLSTM-CRF。评价指标主要集中在精确度、召回率和 F1 分数上，具体的对比实验结果如表 3 所列。

表 3　对比实验结果

Table 3　Comparison of experimental results

（%）

模型	*Precision*	*Recall*	*F*1
BERT	72.63	75.31	73.94
BERT-Softmax	65.87	77.02	71.00
BERT-CRF	75.23	78.80	76.97
BERT-BiGRU-CRF	76.33	79.19	77.73
BERT-BiLSTM-CRF	75.72	77.25	76.47
Roberta-BiLSTM-CRF	77.22	78.45	77.83
Roberta-DFF-BiLSTM-MHA-CRF	**78.21**	**81.53**	**79.84**

实验结果表明，Roberta-DFF-BiLSTM-MHA-CRF 模型在精确度、召回率和 F1 分数这 3 个关键指标上均表现优异。此外，BERT-BiGRU-CRF 模型相较于原始的 BERT 模型，在所有评估指标上都有所提升；BER-BiLSTM-CRF 模型的表现与 BERT-BiGRU-CRF 相似，但在召回率上略低。这些结果揭示了从 BERT 到其与其他模型结合的架构改进，有助于模型更好地捕捉文本中的长短期依赖关系。特别是 BiLSTM 和 BiGRU 这样的双向模型，通过双向的信息流动，能够更全面地理解文本上下文，从而提升模型的整体性能。Roberta-BiLSTM-CRF 模型虽然在召回率上略低于 BERT-BiGRU-CRF，但在精确率上表现更优。这可能是因为 Roberta 在 BERT 的基础上进行了进一步的优化，例如使用了更大的训练数据集和更长的训练序列，这些改进有助于模型更深入地理解文本。在 Boberta-BiLSTM-ATT-CRF 模型中，注意力机制的引入进一步提升了模型对文本中关键信息的捕捉能力。

为探究注意力机制的不同头数对模型性能的影响，选取一组注意力头数(2，4，16，32)进行实验，如图 5 所示。

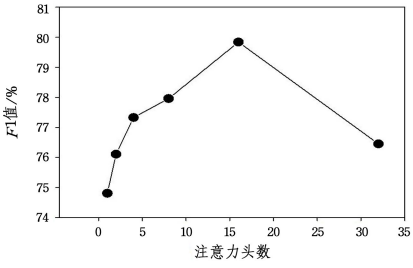

图 5　不同注意力头数对模型的影响

Fig. 5　Impact of different numbers of attention heads on the model

从图 5 中可知，当注意力头数从 1 增加到 4 时，*F*1 值逐步升高。当注意力头数继续增加到 8 时，*F*1 值保持不变。当注意力头数为 16 时，*F*1 值达到最高，为 0.79。然而，当头数增加到 32 时，*F*1 值下降。这意味着在该模型的配置和数据集上，16 头能够最好地捕捉到数据中的复杂特征和关系，从而实现最佳的性能。

批训练大小 batch_size 和学习率的选择通常需要通过实验调整来确定，以找到训练效率和模型性能之间的最佳平衡点。批训练大小与 F1 得分的关系如图 6 所示，学习率与 F1 得分的关系如图 7 所示。可以看出，在批训练大小为 4 时，*F*1 得分达到峰值；学习率为 3×10^{-5} 时，*F*1 得分达到峰值。

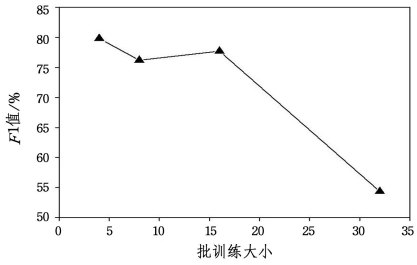

图 6　批训练大小对模型的影响

Fig. 6　Effect of batch training size on model

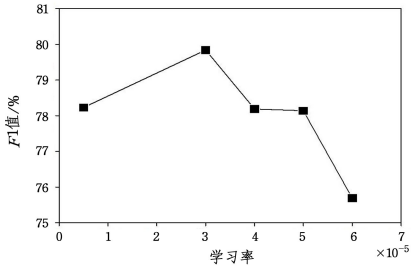

图 7　学习率大小对模型的影响

Fig. 7　Effect of learning rate on model

4.5　消融实验

为了深入分析模型中各个模块对命名实体识别任务性能的具体影响，进行了一系列的消融实验。通过将 Roberta-DFF-BiLSTM-MHA-CRF 模型的各个部分逐步移除，来验证各部分对模型的重要性，结果如表 4 所列。

表 4　消融实验结果

Table 4　Results of ablation experiment

（%）

模型	*Precision*	*Recall*	*F*1
Roberta-DFF-BiLSTM-MHA-CRF	**78.21**	**81.53**	**79.84**
－MHA	77.02	79.46	77.59
－BiLSTM	76.84	80.66	77.81
－CRF	69.14	78.93	72.56
－DFF	74.52	78.26	75.59
－(BiLSTM，MHA，CRF，DFF)	56.97	72.66	62.92

从表 4 中看出，当移除多头注意力机制(MHA)后，模型的 *Precision*，*Recall* 和 *F*1 分数均有所下降，说明 MHA 对于提高模型性能有正面作用。移除 DFF 后，Precision 和 *F*1 分数均有所下降，*F*1 分数从 79.84%降至 75.59%，说明 DFF 在处理输入特征和增强模型表达能力方面起到了一定作用。移除 BiLSTM 后，虽然 *Precision* 略有下降，但 *Recall* 却有所上升，可能是因为 BiLSTM 有助于捕捉序列中的长期依赖关系。CRF 层的移除导致 *Precision* 显著下降，而 *Recall* 则有轻微的提升，这可能是由于 CRF 能够有效地优化标签序列的预测，确保标签之间的转换更加合理。当同时移除 BiLSTM，MHA，CRF 和 DFF 时，模型的性能大幅下降，尤其是 *Precision* 和 *F*1 分数分别降到了 56.97%和 62.92%。这表明这些组件共同作用，对模型的最终性能至关重要。

由此可得，这些组件各自在不同的方面对模型的性能有着积极的影响，其中 CRF 和 MHA 对提高模型的 *Precision* 和 *F*1 分数尤为重要，而 BiLSTM 在捕捉序列信息方面发挥了作用。

综上分析，本文提出的 Roberta-DFF-BiLSTM-MHA-CRF 模型在招标文件数据集上的表现较好。这一模型的优

势在于其综合了文本的特征，并深入分析了文本上下文，这些因素共同促进了模型性能的提升。

结束语 本文针对招标领域的命名实体识别任务，构建了招标文件数据集，并提出了 Roberta-BiLSTM-CRF 的命名实体识别方法。通过融入多头自注意力机制，显著提升了对长跨度实体的识别准确度，并通过整合领域特定的专业术语，有效解决了术语边界模糊的问题，进而增强了对招标文本中实体的识别能力。在真实招标文件数据集上的实验结果证明了该方法的有效性，相较于 BERT-BiLSTM-CRF，本文提出的模型在 Precision 上提升了 2.49 个百分点，在 Recall 上提升了 4.28 个百分点，在 F1 上提升了 3.37 个百分点。

RoBERTa-DFF-BiLSTM-MHA-CRF 模型在招标文件命名实体识别任务中表现出了显著的性能提升，通过整合多头注意力机制和词典特征融合，不仅提高了模型对长跨度实体的识别能力，还增强了对专业术语边界的识别能力。尽管模型在招标文件数据集上表现优异，但其一般性可能受到限制。未来的工作将探索如何将这些方法应用到更广泛的领域和数据集上，以及如何在标注数据受限的情况下提升模型性能。

参 考 文 献

[1] SHI B. Problems in the preparation of bidding documents and rationalization suggestions [J]. China Tendering, 2023(8): 137-138.

[2] MCCALLUM A, FREITAG D, PEREIRA F. Maximum entropyMarkov models for information extraction and segmentation [C]// Proceedings of 17th International Conference on Machine Learning. 2000: 591-598.

[3] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditionalrandom fields: probabilistic models for segmenting andlabeling sequence data[C]// Proceedings of 17th International Conference on Machine Learning. San Francisco: MorganKaufmann Publishers, 2001: 282-289.

[4] YU J D, FAN X Z, YIN J H. Application of hidden Markov model in natural language processing[J]. Computer Engineering and Design, 2007, 28(22): 5514-5516.

[5] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[C]// Proceedings of the Association for Computational Linguistics (ACL). 2014: 655-665.

[6] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]// Proceedings of NAACL-HLT. 2016: 260-270.

[7] ZHU X, SOBIHANI P, GUO H. Long short-term memory over recursive structures[C]// Proceedings of the 32nd International Conference on Machine Learning(LCML-15). 2015: 1604-1612.

[8] HUANG Z, WEI X, KAI Y. Bidirectional LSTM-CRF modelsfor sequence tagging[J]. arXiv: 1508.01991, 2015.

[9] ZHANG S F, WEN L Y, BIAN X, et al. Oc-clusion-aware r-cnn: Detecting pedestrians in a crowd[J]. The European Conference on Computer Vision(ECCV), 2018, 11207: 657-674.

[10] LIU W, LIAO S C, HU W D, et al. Learningefficient single-stage pedestrian detectors by asymptoticlocalization fitting[C]// Computer Vision—ECCV 2018. 2018: 643-659.

[11] ZHANG S S, BENENSON R, SCHIELE B. Citypersons: A diverse dataset for pedestrian detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4457-4465.

[12] LUO B, ZHANG X F, DUAN L, et al. Military Named Entity Recognition Based on RoBERTa-Span-Attack Label Pointer Network[J]. Journal of Naval University of Engineering, 2024, 36(1): 76-82, 93.

[13] LI J H, XIONG W, GONG K, et al. Research on Entity Recognition of Power Equipment Defects Integrating BERT－WWM and Attention Mechanism[J]. Journal of Electric Power, 2024, 39(2): 126-135.

[14] ZHANG Y C, YANG Y, JIANG R, et al. A Business Entity Recognition Model Based on BiLSTM-CRF[J]. Computer Engineering, 2019, 45(5): 308-314.

[15] MI J X, XIE H W. Research and Application of Named Entity Recognition for Bidding Materials[J]. Computer Engineering and Applications, 2023, 59(2): 314-320.

[16] AEJAS B, BELHI A, ZHANG H, et al. Deep learning-based automatic analysis of legal contracts: a named entity recognition benchmark [J]. Neural Computing and Applications, 2024, 36(23): 14465-14481.

[17] AHMET T, METIN T. Enhanced Named Entity Recognition algorithm for financial document verification [J]. The Journal of Supercomputing, 2023, 79(17): 19431-19451.

[18] MA J, YU Y. Automatic Extraction Method for Key Information in Logistics Bidding Documents[J]. Computer and Digital Engineering, 2024, 52(5): 1400-1405.

[19] HEIM G. Named entity recognition indigitalen sammlungenein werkstattbericht aus der badischen landesbibliothek[J]. Bibliotheksdienst, 2023, 57(6): 364-375.

[20] PEI D, JING M, LIU H, et al. A fast RetinaNet fusionframework for multi-spectral pedestrian detection[EB/OL]. https://doi.org/10.1016/j.infrared.2019.103178.

[21] MAO H L, AIZIERGUL I, CHEN D G. Named Entity Recognition in Power Grid Dispatching Domain Based on Multi-Head Attention[J]. Computer Technology and Development, 2023, 33(2): 181-186, 194.

[22] LUO X, XIA X Y, AN Y, et al. Chinese Clinical Entity Recognition Combining Multi-Head Self-Attention Mechanism and BiLSTM-CRF[J]. Journal of Hunan University (Natural Sciences Edition), 2021, 48(4): 45-55.

[23] LI B, WANG H C. Implementation and Application of a Chinese Grammar Error Diagnosis System Based on CRF [J]. Computer Science, 2024, 51(S1): 1141-1146.

YANG Hua, born in 1997, postgraduate. Her main research interests include natural language processing and deep learning.

WANG Baohui, born in 1973, senior engineer, master supervisor. His main research interests include software architecture, big data and artificial intelligence.