石油科学通报
Petroleum Science Bulletin

融合旋转位置编码与掩码条件随机场的钻井工程命名实体智能识别方法

曹倩雯[1]，李维[1]，林伯韬[1*]，金衍[1,2]，韩雪银[1,3]，张家豪[1]

1 中国石油大学(北京)人工智能学院，北京 102249
2 中国石油大学(北京)石油工程学院，北京 102249
3 中海油能源发展有限公司工程技术分公司，天津 300452
* 通信作者, linbotao@cup.edu.cn

收稿日期: 2024−07−24；修回日期: 2024−09−12
国家自然基金项目(No. 62402526) 和中国石油大学(北京) 科研启动基金项目(2462024BJRC013) 联合资助

摘要　钻井工程报告记录了油气藏的地质信息以及钻井工程的参数，自动提取报告中的非结构化信息能够显著提高数据入湖的效率，从而实现高效数据管理。然而，这类报告通常具有特定领域的特征，且结构和语言的多样性给命名实体的准确识别带来了诸多挑战。目前，命名实体识别常用的深度神经网络模型通常基于小规模标注数据集进行训练或微调，导致两方面问题。首先，缺乏大规模的标注语料库，限制了训练样本的多样性，进而导致模型在面对新数据或未见过的数据时表现不佳，降低了模型在不同类型数据上的泛化能力。其次，现有模型缺乏针对长距离上下文的文本建模能力，由于相关实体可能分散在钻井工程报告内较长的文本段落中，这类方法难以有效捕获和识别复杂文档中命名实体的关系。为了解决上述问题，本文提出了一种融合旋转位置编码和掩码条件随机场的钻井工程命名实体智能识别方法。该方法基于Transformer编码器、双向长短期记忆网络(BiLSTM)和条件随机场(CRF)架构。Transformer编码器利用预训练语言模型提供丰富的上下文语义表示，BiLSTM捕捉序列依赖性，而CRF则用于序列标注。此外，通过设计掩码建模机制改进了传统的CRF，限制了倒置序列的生成，提高了序列标注次序的一致性。旋转位置编码的集成进一步增强了模型对文本中相对位置信息的感知，促进模型捕捉远距离单词之间的依赖关系，从而提高识别跨越较大上下文范围的命名实体的能力。除了模型改进之外，本文还通过构建领域特定的命名实体语料库来解决训练数据不足的问题。该语料库包括 12 类实体的标注，覆盖了共 20 727 个实体标签，分布于 4 000 个文本段落中，为模型提供了更多样化的训练样本，帮助提高模型的泛化能力。实验结果表明，本文提出的模型在测试集上的$F1$值为 86.49，相较于之前的最优模型提高了 2.65，在长尾分布的实体识别上的性能也显著提高。该方法不仅扩展了命名实体识别在钻井工程中的应用，还能够为工程师提供高效的信息提取工具，加速钻井数据的分析，提高钻井操作管理的效率，并增强数据入湖的效率，从而对钻井项目的决策过程带来积极影响。

关键词　命名实体识别；钻井工程；Transformer编码器；自然语言处理；深度学习

中图分类号：TP37；TE22

引用格式：曹倩雯, 李维, 林伯韬, 金衍, 韩雪银, 张家豪. 融合旋转位置编码与掩码条件随机场的钻井工程命名实体智能识别方法. 石油科学通报, 2024, 09(05): 750−763
CAO Qianwen, LI Wei, LIN Botao, JIN Yan, HAN Xueyin, ZHANG Jiahao1. Intelligent named entities recognition for drilling engineering by integrating rotational position embedding and masked conditional random fields. Petroleum Science Bulletin, 2024, 09(05): 750-763. doi: 10.3969/j.issn.2096-1693.2024.05.057

Intelligent named entities recognition for drilling engineering by integrating rotational position embedding and masked conditional random fields

CAO Qianwen[1], LI Wei[1], LIN Botao[1], JIN Yan[1,2], HAN Xueyin[1,3], ZHANG Jiahao[1]

1 College of Artificial Intelligence, China University of Petroleum- Beijing, Beijing 102249, China
2 College of Petroleum Engineering, China University of Petroleum- Beijing, Beijing 102249, China
3 China National Offshore Oil Corporation Energy Development Co., Ltd. Engineering Technology Branch, Tianjin 300452, China

Received: 2024−07−24; Revised: 2024−09−12

Abstract Drilling engineering reports record geological information about oil and gas reservoirs as well as various drilling engineering parameters. The automatic extraction of unstructured information from these reports can significantly improve the efficiency of data integration into data lakes, thereby enabling more efficient data management. However, these reports typically have domain-specific characteristics, and the diversity of their structure and language presents considerable challenges for accurate named entity recognition (NER). Currently, deep neural network models commonly used for NER are typically trained or fine-tuned on small-scale annotated datasets, leading to two main issues. First, the lack of large-scale annotated corpora limits the diversity of training samples, which in turn causes poor performance when the model encounters new or unseen data, decreasing the model's generalization ability across different types of data. Second, existing models lack the ability to effectively model long-distance contextual information in texts. Since relevant entities may be scattered across long text segments in drilling engineering reports, these methods often struggle to capture and recognize relationships between named entities in complex documents. To address the aforementioned issues, this paper proposes an intelligent method for named entity recognition in drilling engineering that integrates rotational position embedding and masked conditional random fields. The proposed method is based on a Transformer encoder, a bidirectional long short-term memory network (BiLSTM), and a conditional random field (CRF) architecture. The Transformer encoder leverages pre-trained language models to provide rich contextual semantic representations, BiLSTM captures sequential dependencies, and CRF is used for sequence labeling. Moreover, the traditional CRF is improved by designing a masked modeling mechanism, which restricts the generation of inverted sequences, thereby enhancing the consistency of sequence labeling order. The integration of rotational position embedding further enhances the model's awareness of relative positional information in the text, allowing the model to better capture dependencies between distant words. This improves the model's ability to recognize named entities spread across larger contextual ranges. In addition to model improvements, this paper also addresses the issue of insufficient training data by constructing a domain-specific named entity corpus. This corpus includes annotations for 12 categories of entities, covering a total of 20,727 entity labels across 4,000 text segments. This enriched dataset provides more diverse training samples, which helps improve the model's generalization ability. Experimental results show that the proposed model achieves an $F1$ score of 86.49 on the test set, representing an improvement of 2.65 percentage points over the previous best-performing model. Furthermore, the model demonstrates significant improvements in recognizing entities with long-tail distributions, which are often underrepresented in typical training datasets. This method not only expands the application of named entity recognition in the field of drilling engineering but also provides engineers with an efficient tool for extracting critical information. By accelerating the analysis of drilling data, it improves the efficiency of drilling operations management and enhances data lake integration, ultimately bringing positive impacts to the decision-making process in drilling projects.

Keywords named entity recognition; drilling engineering; transformer encoder; natural language processing; deep learning

doi: 10.3969/j.issn.2096-1693.2024.05.057

0 引言

钻井工程报告包含地质与工程设计等关键信息，报告数据以半结构化与非结构化形式存在，多样且分散，专业性强，提取难度大。快速准确分析这些信息对提高钻井作业管理效率、缩短钻井周期及确保作业的安全环保至关重要[1−3]。尽管报告编写有规范可循，但实际操作中文本结构差异大、传统信息抽取方法效率低下且适应性差。因此，研发高效精准的钻井工程

命名实体识别技术，自动化挖掘专业领域的实体信息，成为钻井工程数字化亟待解决的关键问题[4]。

鉴于工程报告内容的复杂性与多样性，以及信息抽取技术面临诸多局限，促使学术界与工业界不断探索更加高效、智能的解决方案。苏庆林等[5]于 2005 年率先为油田科技情报系统设计了非结构化数据库方案，通过二维表形式明确定义数据集合，实现了对文章标题、正文等关键信息的全文检索，为后续研究奠定了基础。文必龙等[6]设计了GATE框架，通过语法分析和规则抽取取得了一定成效；但鉴于钻井工程报告的语言复杂性，其精确度仍有待提升。李云静等[7]采用词典与结构分析结合的方法，结合决策树模型，有效解决了显性代词指代的消解难题，提升了信息抽取的精度。近年来，深度学习技术的迅猛发展为钻井工程领域的命名实体识别研究提供了新途径。Hoffimann等[8]通过结合并充分发挥卷积神经网络(CNN)与长短期记忆网络(LSTM)两种网络的深度学习能力，实现了对钻井报告中作业指令的分类与预测，取得显著效果。钟原等[9]采用双向长短期记忆网络(Bi−LSTM)与条件随机场(CRF)构建命名实体识别模型，通过预训练词向量和条件随机场来学习标签的约束条件，大幅增强了模型对上下文语义的理解能力。随着知识图谱理论的不断发展，其在钻井工程领域的应用也日益广泛。Yuan等[10]为石油和天然气行业设计了数据语义标准化模型，并结合国际标准构建了行业参考词汇标准化模型，为构建行业标准化的知识图谱提供了重要支撑。Lee等[11]基于上下文化框架，采用RoBERTa预训练模型，成功从非结构化数据中提取命名实体信息，提升了命名实体识别的准确性和实用性。高国忠等[12]在油气领域中首次引入了BERT−BiLSTM−CRF实体识别架构，为实体识别领域研究提供了重要的参照和对比基础；由此，该架构亦用作本文的重要对比方法。

现有命名实体识别的研究在钻井工程领域已取得阶段成果，但现有的传统规则模型灵活性、鲁棒性不足，而深度学习模型处理大规模数据时的通用性和可扩展性则亟待提升[13−15]。由此，本文结合预训练语言模型(PLMs)与旋转位置编码(RoPE)，利用双向长短期记忆网络增强语境理解能力，并结合条件随机场开展实体标注，精准提取复杂实体语义。同时，引入掩码条件随机场(Masked CRF)优化模型，限制倒置序列，提升识别准确率，从而实现钻井工程报告非结构化数据的自动提取与分析。本文所提方法首先构建大规模垂直领域数据库，再通过深度神经网络模型高效提取专业文本中的关键信息，加速钻井数据分析的自动化，提升钻井工程管理及作业效率。同时，其能够有效管理海量报告信息，实现报告数据的提取与分析，提升信息管理效率与优化管理流程。

1 原理与方法

1.1 Transformer

Transformer模型作为自然语言处理领域的一次里程碑式创新，极大地推动了序列学习任务的发展。在Vaswani等[13]于 2017 年发表的著名论文《Attention is All You Need》中，Transformer模型首次被提出，旨在解决传统循环神经网络(RNN)和长短期记忆网络在处理序列学习任务时，并行化计算的困难及捕捉长距离依赖关系的不足[17]。Transformer模型的整体架构如图 1 所示，主要包含编码器(Encoder)和解码器(Decoder)两部分。Encoder负责捕获输入序列的深层次特征，而Decoder则基于这些特征生成对应的输出序列。

在模型处理过程中，输入的文本序列首先进入嵌入层(Input Embedding)开展分词和词嵌入处理，转化为张量$X_{\text{input}} \in \mathbb{R}^{b\times l\times d}$，其中$b$表示批次大小(Batch Size)，即一次输入模型的样本数。l表示序列长度(Sequence Length)，即每个样本中包含的字符或词的数量，d表示词嵌入的维度(Embedding Dimension)，即每个字符或词被嵌入到的向量空间的维度。针对每个词向量，采用正弦和余弦函数的线性组合来生成其位置编码。具体而言，对于偶数位置，引入正弦函数变量；而对于奇数位置，则采用余弦函数变量。据此，区分序列中不同位置的信息，有效捕捉到序列中每个词的位置信息PE，进而提升模型的性能。令字符位置为pos；位置编码中的维度索引为i，计算PE的方法可表示为：

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{\text{model}}}\right) \tag{1}$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{\text{model}}}\right) \tag{2}$$

将X_{input}与位置编码相加得到$X_{\text{embedding}}$作为Encoder的输入X。在其结构中，每一层级都融合了自注意力机制以及前馈神经网络，从而能够精准捕获并理解输入序列所蕴含的上下文信息。在自注意力机制中，通过线性变换将$X_{\text{embedding}}$映射到查询Q、键K和值V表示。令线性变换的权重矩阵为XW_Q、XW_K和XW_V，计算方法可表示为：

$$Q,K,V = XW_Q, XW_K, XW_V \tag{3}$$

得到查询Q、键K和值V之后，计算注意力分数并进行

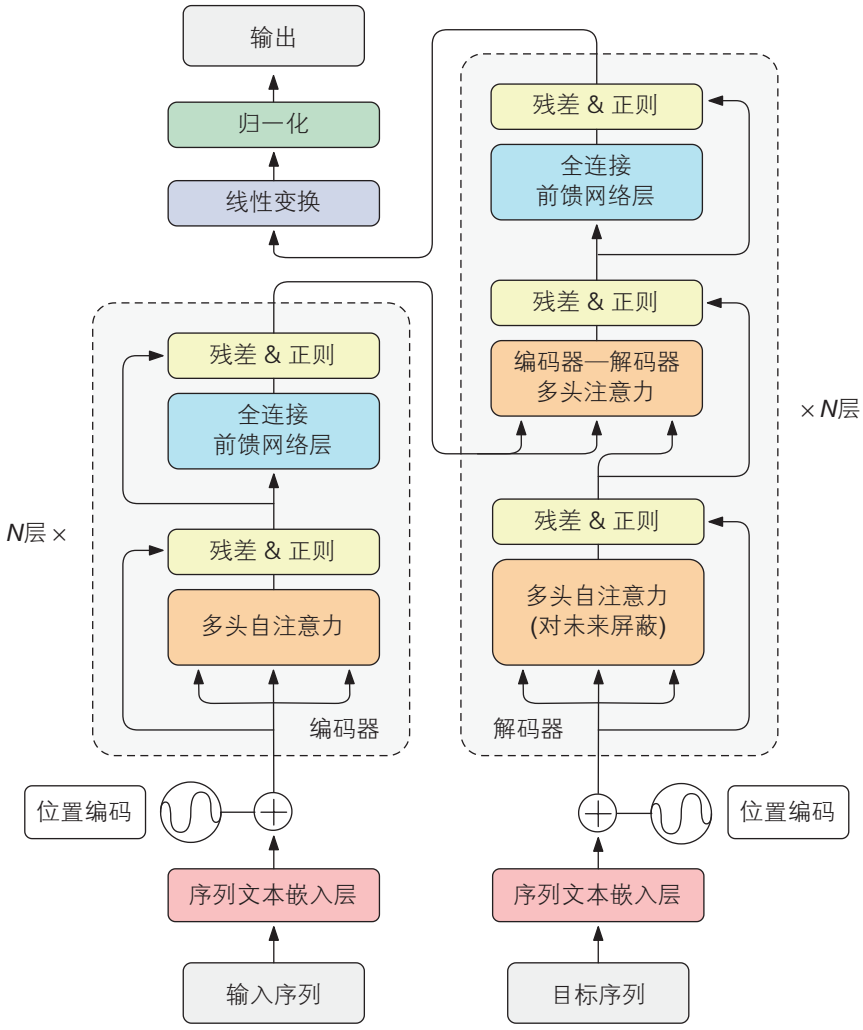

图 1　Transformer模型结构图(改自Vaswani等[9])

Fig. 1　Architecture of the Transformer model (modified from Vaswani et al.[9])

加权。假设有h个注意力头(即多头注意力的数量)，每个头都有查询Q、键K和值V的线性变换权重矩阵，可学习不同的注意力表示。最后，将多个注意力头的输出拼接在一起，并经过线性变换获取最终的注意力输出。多头自注意力计算可表示为：

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{4}$$

$$MultiHead\left(Q,K,V\right)=\left[Attention_1,Attention_2,\ldots,Attention_n\right]\cdot W_O \tag{5}$$

之后，将得到的多头注意力结果进行残差连接与归一化处理，送入全连接前馈网络层(FFN)中。同时引入非线性变换和特征交互，从而增加模型的表达能力和学习能力。令权重矩阵为W_1与W_2，偏置向量为b_1与b_2，使用$ReLU$激活函数，计算方法可表示为：

$$FFN\left(X\right)=ReLU\left(X\cdot W_1+b_1\right)\cdot W_2+b_2 \tag{6}$$

随后基于结果，再次开展一次残差连接与归一化处理，之后将其传递给Decoder。Decoder模块在处理过程中同样采用了自注意力机制、前馈神经网络以及增设的编码器—解码器注意力(Encoder-Decoder Attention)子层。解码器首先通过自注意力机制处理自身输入序列，然后利用编码器—解码器注意力子层关注编码器输出的序列特征，最后将处理结果送入线性变换层和softmax层，生成针对输入序列中各个位置的概率分布。

Transformer模型通过自注意力模块的结构，显著增强了其并行计算的能力，从而使其能够迅速且高效地应对庞大的数据集。此外，该模型还具备出色的长距离依赖捕获能力，有助于更深入地理解输入序列的上下文信息，进而提升模型的整体性能。

1.2 预训练语言模型

随着深度学习技术的飞速发展，预训练语言模型在诸多自然语言处理任务中展现出显著的优势。特别是在命名实体识别领域，BERT模型凭借其独特的设计，极大地提升了模型的性能。BERT由谷歌于2018年提出[18]。其通过引入双向Transformer编码器，并如图2所示，融合字符嵌入(Token Embeddings)、分句嵌入(Segement Embeddings)以及位置嵌入(Position Embeddings)，捕捉上下文全局语义依赖关系，突破传统模型在文本上下文信息处理方面的局限，为钻井工程的命名实体识别任务提供了新的解决思路。

1.3 双向长短期记忆网络

传统神经网络模型在处理时间序列数据时，往往忽略了序列的双向依赖关系，在一定程度上限制了其性能。双向长短期记忆网络(BiLSTM)起源于长短期记忆网络LSTM[19]，通过同时考虑序列的正向和反向信息，实现了对序列的双向处理[20]。

双向LSTM包含两个主要构成部分：前向与后向LSTM[21]，如图3所示。x_i代表输入序列中的令牌(Token)位置，y_i对应输出序列的Token位置。LSTM的独立节点被标记为A和A'，输出y_i由这两个节点的输出合并得到。处理信息时，前向LSTM依据序列的正向次序进行，主要负责记录前面的信息；后向LSTM则逆着序列的方向处理信息，专注于捕捉后续的信息。通过该处理机制，每一个时间步的输出都能综合前后两个方向的信息，从而更全面、更准确地反映出序列的上下文特征。

1.4 条件随机场

条件随机场(CRF)作为序列标注领域的关键技术，由Lafferty等[22]在2001年率先提出。该技术以统计学习理论为基础，专注于处理序列标注任务中上下文关系与全局依赖性的捕捉问题。CRF在自然语言处理诸多方面，如序列标注、文本分段、命名实体识别和词性标注等任务中，均展现了显著的应用效果。

作为一种无向图模型，CRF通过构建由节点和边组成的图结构来反映其内在逻辑。在此结构中，节点代表不同的标签变量，而边则揭示了标签间的依赖关系。当面对一个特定的输入序列时，CRF能够计算出对应的输出标签序列的条件概率[23]。该算法的核心机制在于，CRF通过最大化条件概率来推断最合逻辑和语境的标签序列，进而实现对输入序列的精确标注。

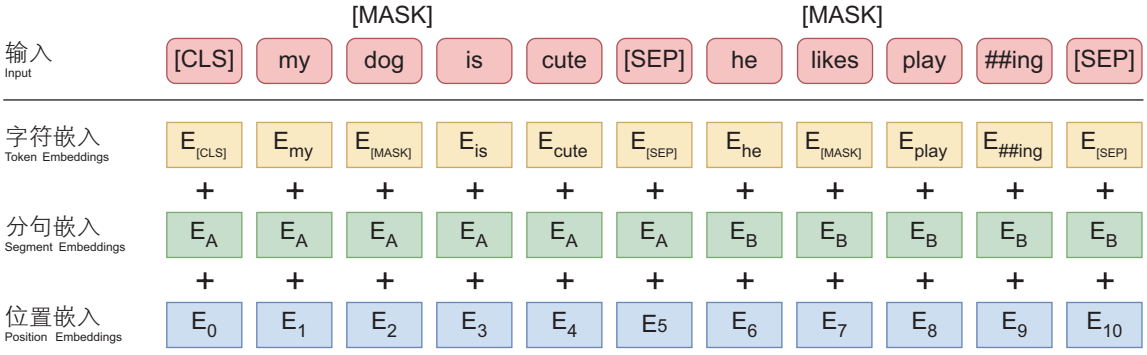

图2 BERT的嵌入层结构图(改自Devlin等[11])

Fig. 2 Embedded layers of the BERT model(modified from Devlin et al.[11])

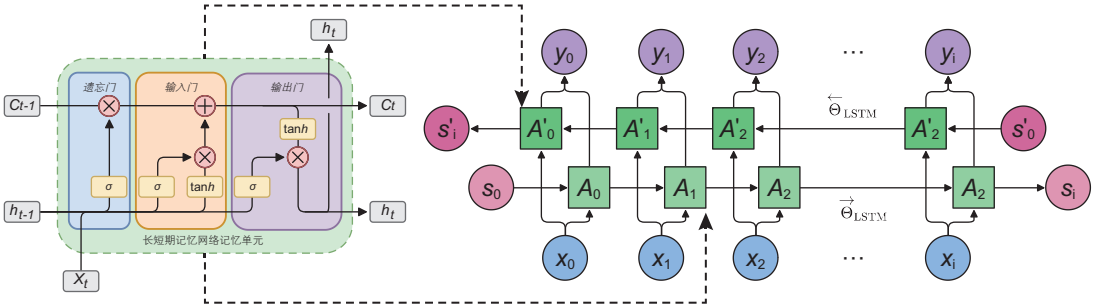

图3 Bi-LSTM模型结构图(改自Jadon等[14])

Fig. 3 Bi-LSTM model architecture (modified from Jadon et al.[14])

2 钻井工程报告命名实体识别方法

2.1 网络结构

为克服传统模型在钻井工程领域命名实体识别中的局限性，提高识别精度，本文提出一种融合旋转位置编码与掩码条件随机场的钻井工程命名实体识别方法(Fusion of Rotational position embedding and Masked conditional random fields)，简称FoRaM。该方法融合了预训练语言模型、双向LSTM和条件随机场的优势，能够实现钻井工程报告文本中的命名实体高效、准确识别。

如图4所示，模型主要包含3个核心层：Transformer层、BiLSTM层 和CRF层。Transformer层作为核心组件，通过捕捉文本上下文信息，学习词汇的丰富表示。随后，BiLSTM层对Transformer层生成的词向量进行特征提取，获取深层次特征表示。最后，CRF层接收BiLSTM层输出的标签分数特征矩阵，利用转移矩阵计算标签间转移概率，确定最佳标注序列，从而精准反映实体边界及其类别信息。系统输出预测

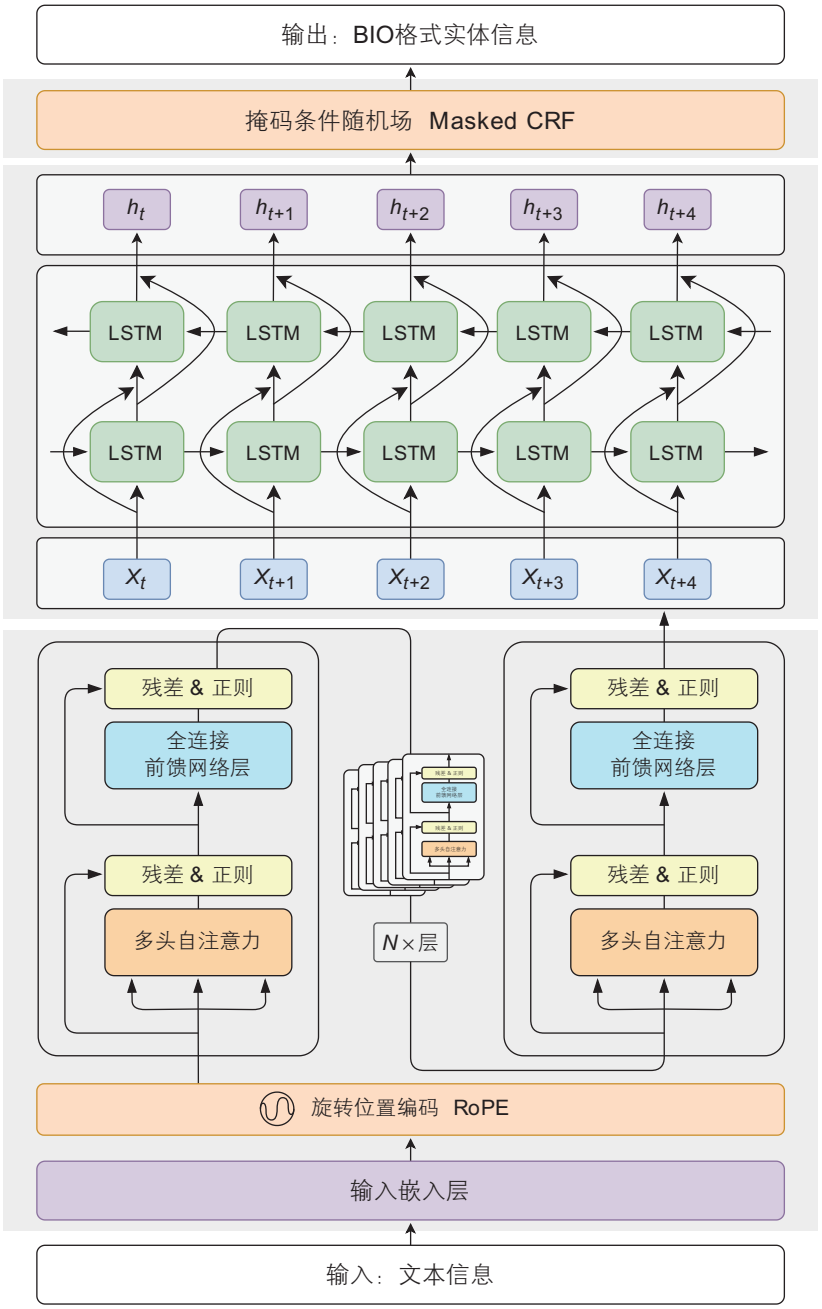

图4 FoRaM模型结构图

Fig. 4 Architecture of the FoRaM model

的BIO标签序列能够精确标注文本中实体的起始、内部及非实体位置。

2.2 旋转位置编码单元(RoPE)

在处理长文本或文本片段时，传统BERT的位置编码方法受限于预设的最大长度，在一定程度上影响了模型对不同长度输入序列的适应能力。为克服该限制，本文引入了苏剑林等[24]提出的旋转式位置编码(Rotary Position Embedding, RoPE)方法优化基线模型。

RoPE技术通过抽取旋转矩阵编码位置信息，显著提升了模型处理不同序列长度时的灵活性。针对钻井报告此种冗长文本或文本片段时，这种灵活性尤为重要。此外，RoPE将绝对位置编码转换为相对位置编码，实现更精确的位置信息嵌入，进而提升模型性能。其核心机制在于利用旋转矩阵编码绝对位置，使相对位置信息能够被自注意力机制捕捉，从而更深入理解文本序列中词汇间的相互关系，提高命名实体识别的准确性。

具体而言，定义旋转矩阵为R_m，对于一个二维向量$\vec{x}$，若将其绕原点旋转m弧度，则会得到变换后的旋转位置编码向量$R_m \cdot \vec{x}$。在位置编码的文本框架内，将查询(Query)向量$\vec{q}$和键(Key)向量$\vec{k}$视作处于二维空间中的向量，则可以通过旋转这些向量对应的位置索引弧度，有效计算出相对位置编码的信息，进一步得到对应的注意力分数score。该旋转操作是RoPE能够实现灵活且精准位置编码的关键所在，计算方法可表示为：

$$R_m = \begin{bmatrix} \cos m & -\sin m \\ \sin m & \cos m \end{bmatrix} \tag{13}$$

$$R_m^{\top} \cdot R_n = R_{n-m} \tag{14}$$

$$\text{score}\left(\vec{q}_m, \vec{k}_n\right) = \left(R_m \cdot \vec{q}_m\right)^{\top} \cdot \left(R_n \cdot \vec{k}_n\right) \tag{15}$$

综上，通过基于旋转的位置编码，使优化后的模型在处理长文本时能够更好地结合相对位置信息，提升上下文信息的准确性，进而增强模型对不同长度输入序列的适应能力和命名实体识别的性能。该方法能够帮助深入理解钻井报告的文本内容、有效执行信息抽取任务。

2.3 掩码条件随机场单元(Masked CRF)

基于深度学习的方法在序列标注任务中展现出了显著的优势，特别是在处理可能产生倒置问题的序列时。Wei等[25]提出了一种创新的掩码条件随机场模型Masked CRF，旨在解决传统条件随机场在处理序列时可能遇到的倒置难题。Masked CRF模型通过在特征函数和转移概率中引入掩码(Mask)机制，实现了对序列中可能存在的倒置实体的有效规避。在模型训练阶段，Masked CRF运用掩码技术将倒置路径的分数设定为负无穷大，从而确保模型在训练过程中能够主动忽略并避免学习到这些倒置的序列模式。这种策略使得模型能够专注于学习和识别符合语法规则的有效序列。

进入解码阶段后，Masked CRF采纳这些约束条件，确保仅输出遵循特定规则的序列。由于倒置路径在训练阶段已被有效屏蔽，解码过程中模型自然不会将这些路径纳入考虑范围，从而彻底消除了产生倒置序列的风险。这种约束的实现依赖于掩码倒置转移矩阵。该矩阵通过设定倒置转移概率为负无穷，使得包含倒置转移的路径在计算路径分数时得分极低，进而在模型评估中被忽略。例如图5所示，输入文本“需要灌注钻井液”到CRF模型中会得到“灌注”是作业内容而“钻井液”出现了倒置序列的问题即“I-物料”开头，违反了B开头的BIO标注规定，并且“灌注钻井液”是一个作业内容的整体，不能将其分开。Masked CRF的预测结果能够完全遵循BIO标注规范，并正确标注了作业内容，见图5。

Masked CRF模型通过实施严格的候选路径约束和掩码技术，成功解决序列标注中的倒置序列问题。它不仅显著提升了模型在复杂序列处理任务中的准确性和效率，还确保了输出序列的语法正确性，进一步增强了模型的可靠性和稳定性。

3 实验评估

3.1 实验数据标注

在钻井工程领域，鉴于目前缺乏大规模开源语料库的现状，首要任务是获取并标注相关数据。本文选取中海油宿州区块、临兴区块以及渤海油田等地共21口井的钻井工程设计报告作为构建语料库的数据来源。在构建语料库的过程中，将原始的钻井报告文件进行文本化处理，即将原本非结构化的数据转换成文本格式。由于钻井工程报告采用*.doc格式，其信息提取受到限制，因此需要将其转换为基于XML的*.docx格式，用以提升文档的阅读体验和信息检索效率。

针对钻井工程设计报告中存在的复杂非结构化信息，如图片、表格等，开展专业信息过滤或提取，确保仅保留与油气钻井工程紧密相关的文本内容。在解析.docx文档时，特殊字符和各类标点符号等会导致模型误读，进而产生错误的解析结果，降低数据的准确性。为此，先剔除数据中的无用特殊字符和标点符号再进行数据清洗，从而保留有价值的文本内容[26]。在

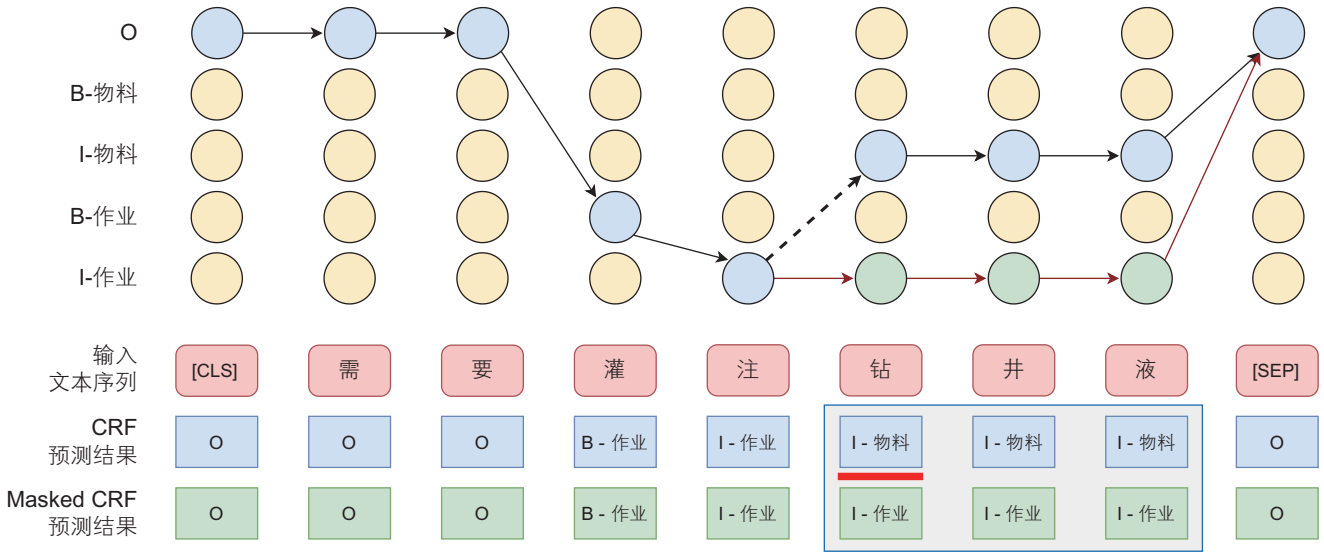

图 5　Masked CRF 预测结果对比图

Fig. 5　Comparison chart of the Masked CRF prediction results

处理文本数据集时，为优化数据质量，首先清除了数据集中的空行、注释及无关紧要的文字，旨在减少噪音数据的影响[27]。其次，对过长句子适当分割，降低计算的复杂度并提升模型的性能，整理获得适用于本实验的待标注语料库。

标注格式的合理选择对模型训练和性能评估具有重要的作用。本文选用BIO标注法，将实体标记划分为起始(Begin，简称B)、内部(Inside，简称I)和非实体(Outside，简称O)3 个部分。具体来说，B用于标识实体的起始位置，I用于表示实体内部的延续部分，而O则代表文本中的非实体成分[28]。以句子“乔布斯是苹果公司的创始人”为例，其中的“乔布斯”会被标注为“B−PER I−PER I−PER”，清晰地指示出这是一个人名实体。针对钻井报告中的文本信息，本文将其划分为 12 种实体标签以便于表示，见表 1，并进行了钻井工程文本标注示意，见图 6。

本文按照 3:1 的比例划分已完成标注的 4000 段文本数据集。此种划分比例不仅有利于保障模型训练的精准度，同时也有助于提高模型的泛化能力。其中，训练集占总数据的 3/4，用于模型参数的训练和调整。测试集占总数据的 1/4，用于最终评估模型的泛化能力。本文统计了 12 种实体标签在数据集中的分布情况，并分别列出了不同实体标签在训练数据集和测试数据集中的数量，以及总实体数量，见表 2。

3.2　实验设置

本文的数值实验采用Python编程语言 3.8.5，基于Pytorch 2.0.0 和Transformer 4.37.2 对本文方法进行程序实现。实验程序部署于配置为Intel(R)Xeon(R)Gold 6133@2.50GHz处理器、32G内存、NVIDIA A100-PCIE 40G显卡、Ubuntu 22.04 操作系统的服务器。在训练主要超参数方面，设置最大序列长度为 512，训练周期数为 32，批量大小为 54，PLMs学习率为 3e-5，CRF学习率为 3e-3，epsilon参数为 1e-8，权重衰减为 0.01，warmup比例为 0.01。

表 1　实体标签设置情况

Table 1　Status of entity labeling

实体类别	实体标签	标记内容
作业相关区域(TA)	B-TA、I-TA	工程技术作业中心
日期(DAT)	B-DAT、I-DAT	2024 年 1 月 1 日
井别(WT)	B-WT、I-WT	探井、开发井
地质层位划分(GSD)	B-GSD、I-GSD	石炭系、上古生界
作业内容(TD)	B-TD、I-TD	钻探、钻进
处理措施(RM)	B-RM、I-RM	调整钻进方向
钻井开次(DO)	B-DO、I-RO	一开、二开
参数信息(PI)	B-PI、I-PI	井深、压力
井名(WN)	B-WN、I-WN	W1-A1 井、W1-A2 井
岩性(LIT)	B-LIT、I-LIT	砂岩、泥岩
作业物料(TM)	B-TM、I-TM	钻头、管柱
工程层位划分(ESD)	B-ESD、I-ESD	井眼、水平井段
非实体字符	O	停用词以及无关字符

实验结果采用了命名实体识别领域广泛应用的三个评价指标：精确率(Precision)、召回率(Recall)和$F1$值($F1$-Score)作为评价指标[29]，其中精确率为正确识别实体与识别结果总数的占比，召回率为正确识别实体与测试集实体总数占比，$F1$ 值为精确率和召回率的

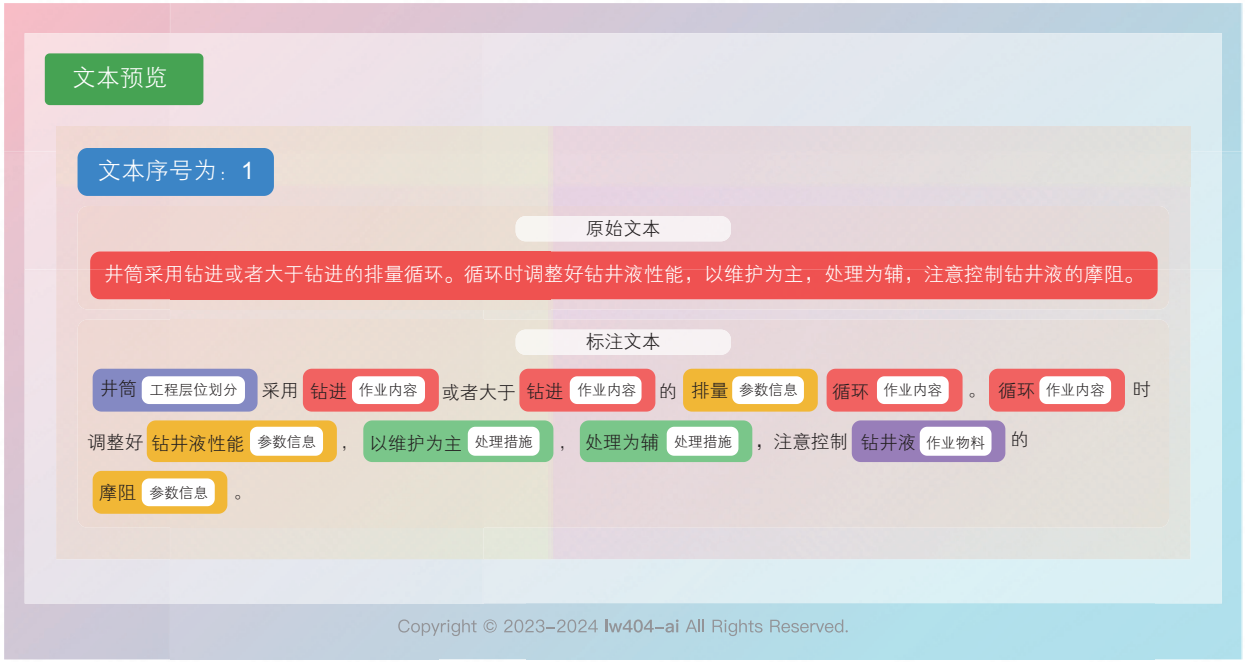

图 6 钻井标注文本示意图

Fig. 6 Schematic diagram of the labeled drilling text

表 2 实验语料中实体标签分布情况

Table 2 Distribution of entity labels in the experimental corpus

实体类别及缩写	训练数据集/个	测试数据集/个	总实体数量/个
日期(DAT)	109	45	154
井名(WN)	445	207	652
钻井开次(DO)	161	58	219
岩性(LIT)	153	84	237
作业内容(TD)	3752	1423	5175
井别(WT)	522	179	701
作业相关区域(TA)	749	231	980
作业物料(TM)	2754	964	3718
工程层位划分(ESD)	1371	466	1837
地质层位划分(GSD)	185	58	243
参数信息(PI)	3948	1309	5257
处理措施(RM)	1191	363	1554
总计	15340	5387	20727

调和平均值，综合考虑了精确率和召回率，体现模型命名实体识别的综合性能。计算方法可表示为：

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{18}$$

其中，真正例(True Positive, TP)表示模型将正类样本正确地预测为正类的数量，假正例(False Positive, FP)表示模型将负类样本错误地预测为正类的数量，真反例(True Negative, TN)表示模型将负类样本正确地预测为负类的数量，假反例(False Negative, FN)表示模型将正类样本错误地预测为负类的数量。

3.3 实验结果分析

本研究针对钻井工程报告中命名实体识别存在的难以满足领域内长段落文本的分析需求，提出了融合旋转位置编码(RoPE)与掩码条件随机场(Masked CRF)的FoRaM实体识别模型。本文使用的主要评估指标包括精确率、召回率和$F1$值，见表 3。实验数据清晰揭示了不同模型间性能的差异。具体而言，在多种复现模型中，CRF模型在精确率上表现相对较弱，仅为55.67%。相较之下，IDCNN−CRF和BiLSTM−CRF模型在精确率上取得了显著提升，分别达到了 71.82%和 74.43%。BERT−BiLSTM−CRF基准模型实现了82.15%的精确率，充分显示深度学习模型在处理序列数据时相较于传统CRF模型的优势。本文所提出的FoRaM模型在所有模型中表现最为突出，精确率高达85.13%，展现出更佳的综合识别效果。

在测试集上，基准模型与基于旋转位置编码模型的精确率对比见图 7。显然，基于RoPE的模型在精确率分布上表现更佳。基线模型的精确率总体分布较低，存在较多低值点。采用RoPE的模型则存在更少的低值点。这一结果不仅验证了FoRaM模型架构的有

效性，同时也证实了RoPE能够提升模型对文本内容相对位置的感知能力。特别是在处理复杂样本时，基于RoPE的模型展现了更为优越的性能，低值点显著减少，有效提高了长文本分析处理能力，对呈长尾分布的实体的识别效果有一定的性能提升。

为进一步探究不同模型在钻井工程领域对各类命名实体的识别能力，对比各模型在识别不同类型命名实体时的性能差异，见图 8。结果表明，本文提出的

表 3　对比实验分析

Table 3　Comparison experiment analysis

类别	模型架构	精确率/%	召回率/%	*F*1 值/%
复现模型	CRF[30]	55.67	48.56	51.74
	IDCNN-CRF[31]	71.82	71.19	71.46
	BiLSTM-CRF[32]	74.43	70.89	72.55
	BERT-BiLSTM-CRF[12]	82.15	81.96	82.04
本文模型	FoRaM	85.13	84.28	84.69

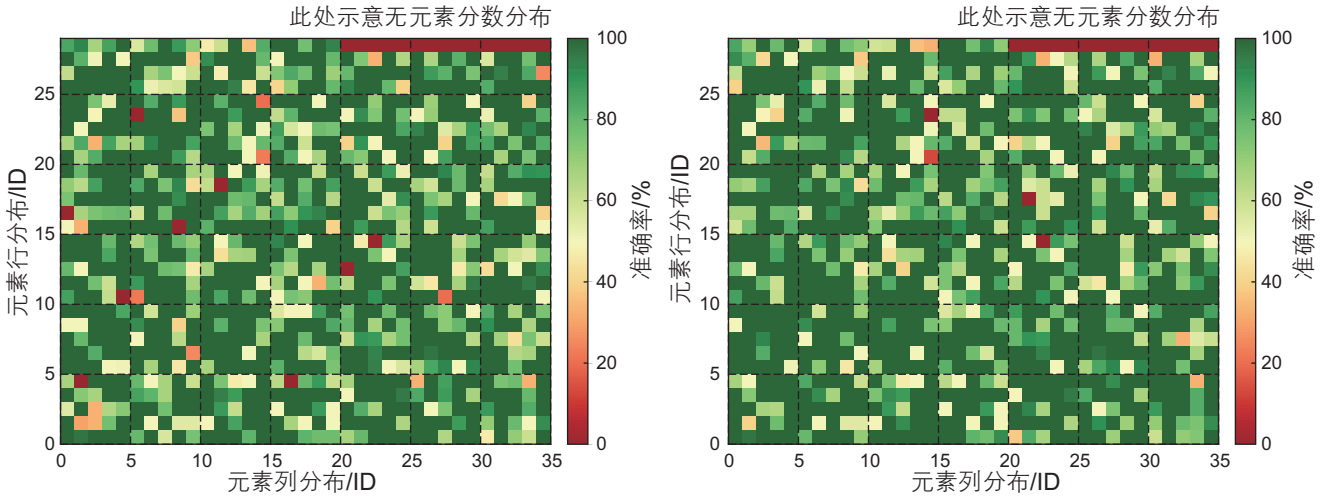

图 7　测试集上基准模型与基于旋转的位置编码模型的精确率对比热力图 (a) 基准模型 (b) 基于旋转位置编码模型

Fig. 7　Heatmap comparing the precision of the baseline and RoPE models on test data set (a) Baseline model (b) Model based on rotary position embedding

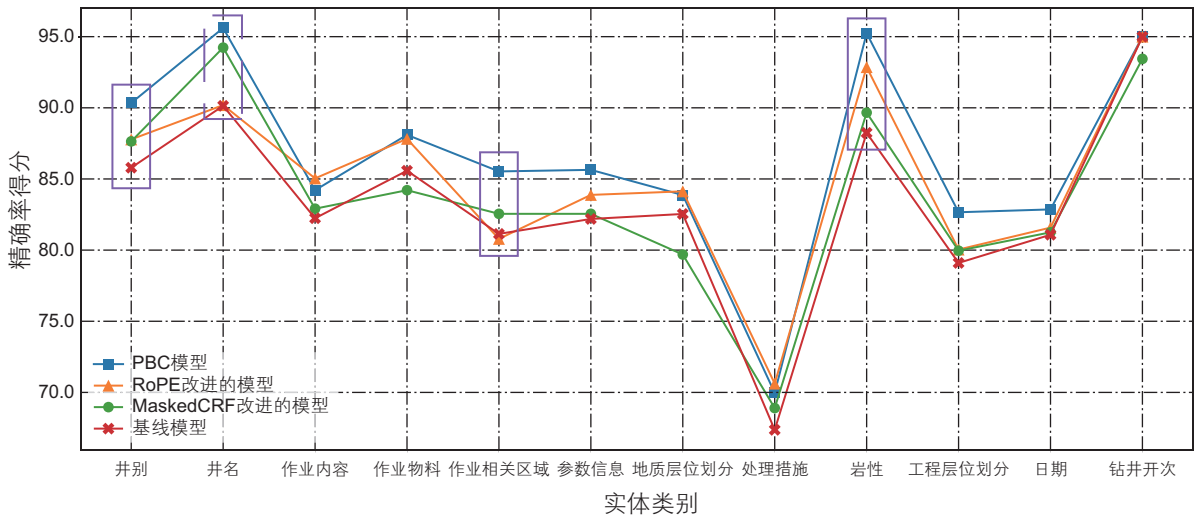

图 8　测试集上不同模型的实体类别得分图

Fig. 8　Category scores of different models on the test data set

FoRaM模型在识别“井别”、“井名”、“作业区域”以及“岩性”等命名实体时均取得了较高得分，验证了FoRaM模型在钻井工程领域的命名实体识别任务中的有效性。与复现模型相比，FoRaM模型在识别“作业内容”、“作业物料”、“参数信息”、“处理措施”、“工程层位划分”等实体多样性更为复杂的类别上得分更高(见图 9)。通过RoPE能够使模型提升在“作业物料”以及“处理措施”等复杂的类别上的性能表现，

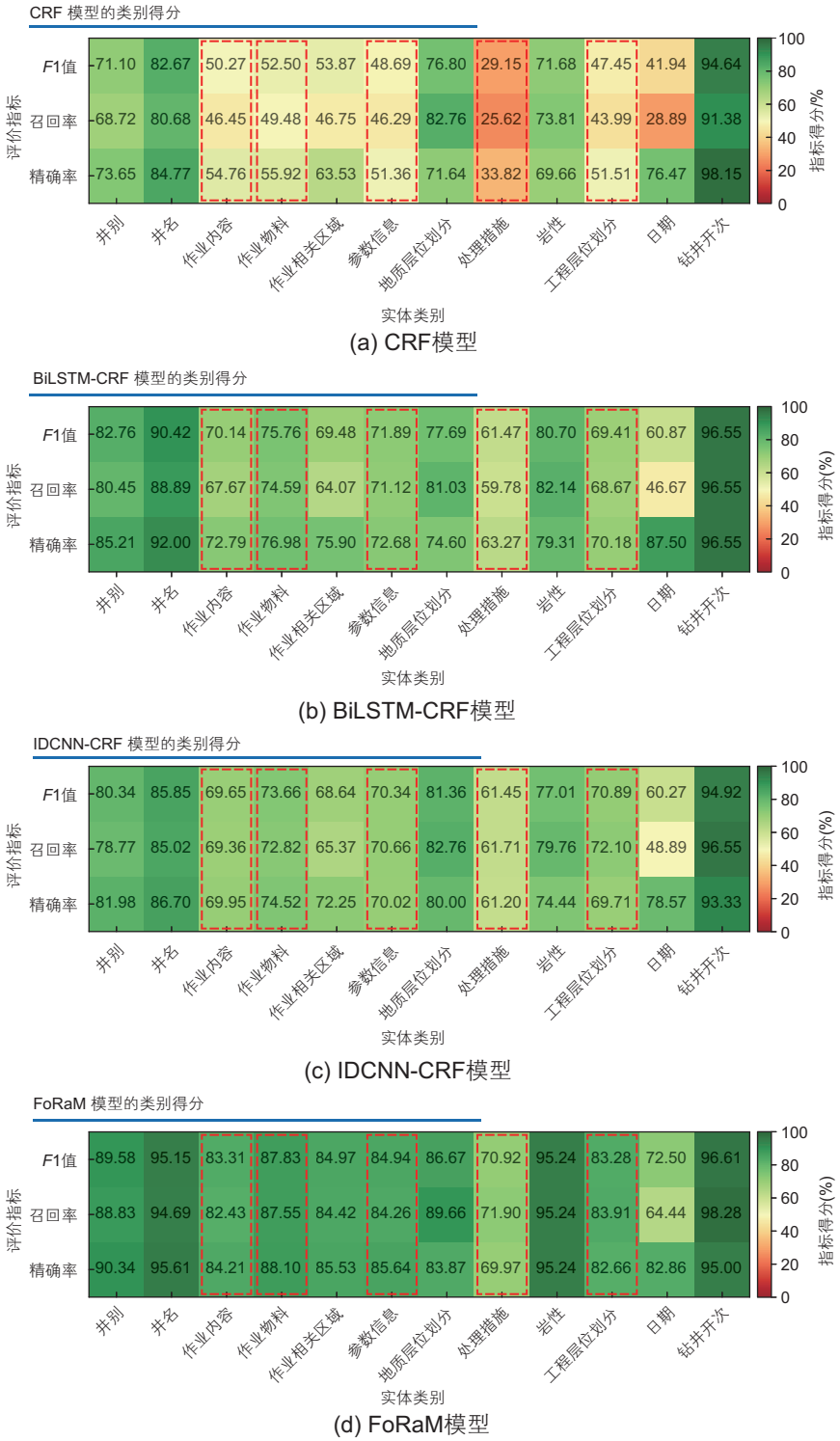

图 9 测试集上不同模型的实体类别得分热力图(a)CRF 模型(b)BiLSTM-CRF 模型(c)IDCNN-CRF 模型(d)FoRaM 模型

Fig. 9 Heatmap of category scores of different models on the test data set (a) CRF model (b) BiLSTM-CRF model (c) IDCNN-CRF model (d) FoRaM model

通过Masked CRF能够确保易出现倒置序列的类别如“井名”等的识别精确率，见表 4。

综上，本文所提钻井工程命名实体识别模型在精确率、召回率和*F*1 值等关键指标上均表现出色，尤其在处理长文本和复杂序列数据时展现了显著的优势。该模型不仅提高了长文本分析处理能力，而且为钻井工程领域的命名实体识别任务提供了新的解决方案。

为探究旋转位置编码(RoPE)与掩码条件随机场(Masked CRF)对本文所提模型FoRaM性能的贡献，进一步展开消融试验分析。实验结果如表 4 所示，相比于基线模型(见表中第一行)，RoPE与Maksed CRF的引入均提高模型性能，并且相比之下，Maksed CRF对于性能提高的贡献更大，验证了所引入两种改进算法单元的有效性。

其次，为探究超参数对于模型性能的影响，本文以训练批次大小(batch size)以及学习率(learning rate)为主要调试对象，批次大小影响模型的泛化能力和训练速度，较小批次有助于降低过拟合风险，而过大则影响泛化性能。调试实验如图 10 所示，为兼顾拟合精度和训练效率，本文所提模型设置批次大小为 36。此外，学习率通常用于控制模型权重更新步长，影响学习速度和收敛稳定性，较小的学习率有利于精确搜索，但训练慢；过大则可能导致训练不稳定。调试实验如图 11 所示，为实现有效收敛并提高训练速度，本文所提模型设置学习率为 3e-5。

表 4　消融实验分析(“+”表示采用该算法单元，“–”表示未采用该算法单元)

Table 4　Ablation experiment analysis (“+” indicates that the algorithm unit is adopted, “–” indicates that the algorithm unit is not adopted)

RoPE	Masked CRF	精确率	召回率	*F*1 值
–	–	82.15	81.96	82.04
+	–	82.60	82.31	82.43
–	+	84.15	84.07	84.09
+	+	85.13	84.28	84.69

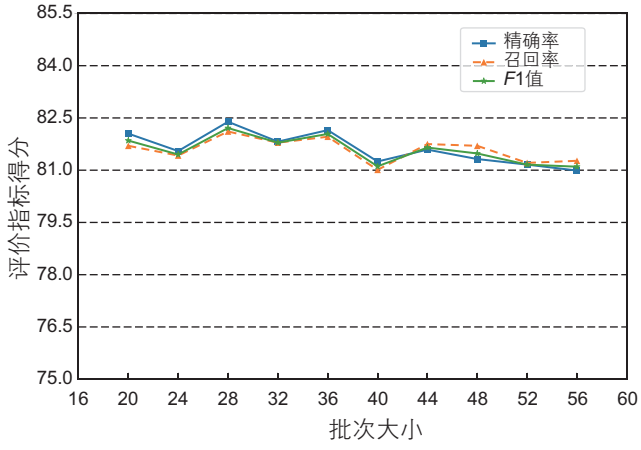

图 10　基线模型性能随批次大小变化图

Fig. 10　Evaluation results varies with the batch size

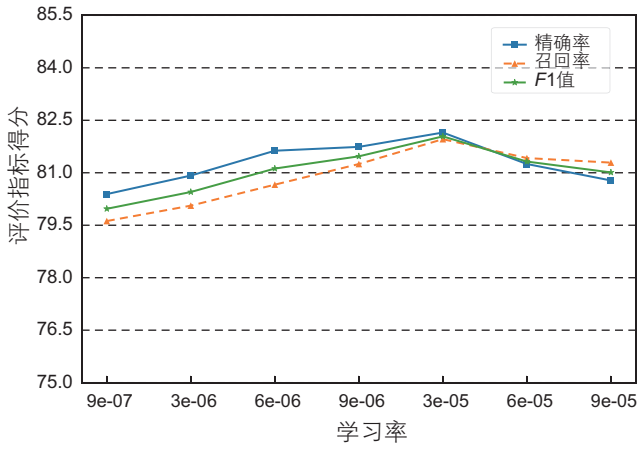

图 11　基线模型性能随学习率变化图

Fig. 11　Evaluation results varies with the learning rate

4　结论

本文基于深度学习和自然语言处理技术，深入探讨了钻井工程报告命名实体识别任务。针对该领域内不同类型命名实体识别效率低下的挑战，提出一种基于Transformer编码器的识别方法，并结合预训练语言模型以捕获钻井工程领域复杂实体的特征，有效提取领域相关的语义信息。为进一步优化模型性能，引入旋转位置编码，以增强模型对序列位置信息的处理能力；同时，为避免倒置序列生成并提升模型性能，通过引入掩码条件随机场，确保模型专注于学习符合语法规则的有效序列，从而提高整体的识别精确率。

实验结果显示，该方法在钻井工程报告关键信息

识别任务中表现良好，相较于传统识别方法，有效提升了准确率，同时在处理速度和稳定性方面亦展现出明细优势。该方法不仅拓宽了命名实体识别技术在钻井工程领域的应用范围，而且为钻井工程师提供了一种高效的非结构化数据抓取及分析工具，能够迅速从海量钻井报告中提取关键信息，进而实现钻井作业信息管理的数字化和自动化，同时也为更广泛领域的实体识别任务提供了有效参考和借鉴。

参考文献

[1] XIANGGUANG Z, RENBIN G, FUGEN S, et al. PetroKG: Construction and application of knowledge graph in upstream area of PetroChina [J]. Journal of Computer Science and Technology, 2020, 35:368−78.

[2] 和婷婷, 张强. 知识图谱在油气勘探开发中的应用现状与发展趋势[J]. Natural Gas Industry, 2024, 1;44(9). [SU Q L, JIN G, CHEN L S. Application status and development trend of knowledge graph in petroleum exploration and development [J]. Natural Gas Industry, 2024, 1;44(9).]

[3] 米石云, 牛敏, 吴珍珍, 等. 全球油气资源信息系统构建与关键技术[J]. 中国石油勘探, 2022, 27(6):145. [MI S Y, NIU M, WU Z Z, et al. Global petroleum resources information system construction and key technology [J]. China Petroleum Exploration, 2005(02): 50.]

[4] HU S, WANG Z, ZHANG B, et al. Data augmentation with knowledge graph-to-text and virtual adversary for specialized-domain chinese NER [C]// Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN), 2024 Jun 30 (pp. 1−8). IEEE.

[5] 苏庆林, 金刚, 陈灵山. 非结构化数据库用于油田科技情报系统 [J]. 油气田地面工程, 2005(02): 50.[SU Q L, JIN G, CHEN L S. Unstructured database for oilfield scientific and technological information system [J]. Oil-Gasfield Surface Engineering, 2005(02): 50.]

[6] 文必龙, 李云静. 基于油田领域本体的信息抽取技术研究 [J]. 计算机技术与发展, 2015, 25(07): 226−229.[WEN B L, LI Y J. Research on information extraction technology based on oilfield domain ontology [J]. Computer Technology and Development, 2015, 25(07): 226−229.]

[7] 李云静. 基于石油领域本体的Web信息抽取技术研究 [D]. 东北石油大学, 2016.[LI Y J. Research on Web information extraction technology based on petroleum domain ontology [D]. Northeast Petroleum University, 2016.]

[8] HOFFIMANN J, MAO Y, WESLEY A, et al. Sequence mining and pattern analysis in drilling reports with deep natural language processing[C]//SPE Annual Technical Conference and Exhibition. SPE, 2018.

[9] 钟原, 刘小溶, 王杰, 等. 基于NER的石油非结构化信息抽取研究[J]. 西南石油大学学报(自然科学版), 2020, 42 (06): 165−173.[ZHONG Y, LIU X R, WANG J, et al. Research on oil unstructured information extraction based on NER [J]. Journal of Southwest Petroleum University (Natural Science Edition), 2020, 42 (06): 165−173.]

[10] YUAN J, LI H. Research on the standardization model of data semantics in the knowledge graph construction of Oil&Gas industry[J]. Computer Standards & Interfaces, 2023, 84: 103705.

[11] LEE M X, MARLOT M. Information retrieval from oil and gas unstructured data with contextualized framework[C]//Third EAGE Digitalization Conference and Exhibition. European Association of Geoscientists & Engineers, 2023, 2023(1): 1−5.

[12] 高国忠, 李宇, 华远鹏, 等. 基于BERT-BiLSTM-CRF模型的油气领域命名实体识别 [J]. 长江大学学报(自然科学版), 2024, 21 (01): 57−65.[GAO G Z, LI Y, HUA Y P, et al. Named entity recognition in the oil and gas field based on the BERT-BiLSTM-CRF model [J]. Journal of Yangtze University (Natural Science Edition), 2024, 21 (01): 57−65.]

[13] 王刘坤, 李功权. 基于 GeoERNIE-BiLSTM-Attention-CRF 模型的地质命名实体识别 [J]. 地质科学, 2023, 58 (3): 1164−1177. [WANG L K, LI G Q. Geological named entity recognition based on GeoERNIE-BiLSTM-Attention-CRF model [J]. Chinese Journal of Geology, 2023, 58 (3): 1164−1177.]

[14] CONSOLI B, SANTOS J, GOMES D, et al. Embeddings for named entity recognition in geoscience portuguese literature [C]// Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 4625−4630.

[15] 宋先知, 姚学喆, 李根生, 等. 基于 LSTM-BP 神经网络的地层孔隙压力计算方法[J]. 石油科学通报，2022, 7(01):12−23. [SONG X Z, YAO X Z, LI G S, et al. A novel method to calculate formation pressure based on the LSTM-BP neural network [J]. Petroleum Science Bulletin, 2022, 7(01):12−23.]

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[17] SOCHER R, LIN C C, MANNING C, et al. Parsing natural scenes and natural language with recursive neural networks[C]//Proceedings of the 28th international conference on machine learning (ICML−11). 2011: 129−136.

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

[19] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735−1780.

[20] 王海涛, 王建华, 邱晨, 等. 基于双向长短期记忆循环神经网络和条件随机场的钻井工况识别方法[J].石油钻采工艺,2023,45(05):540−547+554.[WANG H T, WANG J H, QIU C, et al. Drilling condition identification method based on bidirectional long and short-term memory recurrent neural network and conditional random field [J]. Oil Drilling & Production Technology, 2023, 45(05): 540−547+554.]

[21] JADON S, MILCZED J K, PATANKAR A. Challenges and approaches to time-series forecasting in data center telemetry: A survey[J]. arXiv preprint arXiv:2101.04224, 2021.

[22] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.

[23] 李连兄, 张云天. 基于条件随机场的中文分词技术研究[J].信息技术与信息化,2022(08):116−118+122.[LI L X, ZHANG Y T. Research on Chinese word segmentation technology based on conditional random fields [J]. Information Technology and Informatization, 2022(08): 116−118+122.]

[24] SU J, AHMED M, LU Y, et al. Roformer: enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024, 568: 127063.

[25] WEI T, QI J, HE S, et al. Masked conditional random fields for sequence labeling[J]. arXiv preprint arXiv:2103.10682, 2021.

[26] 李太帆, 王娟, 马良乾, 等. 基于深度学习与规则匹配的Word文档实体识别与属性抽取融合算法及其在油气勘探领域中的应用[J]. 信息与电脑(理论版), 2023, 35(11): 92−96.[LI T F, WANG J, MA L G, et al. A fusion algorithm for word document entity recognition and attribute extraction based on deep learning and rule matching and its application in the field of oil and gas exploration [J]. Information and Computer (Theoretical Edition), 2023, 35(11): 92−96.]

[27] 张超群. 基于知识图谱的钻井作业决策支持系统研究与开发[D]. 西安石油大学, 2022. [ZHANG C Q. Research and development of a decision support system for drilling operations based on knowledge graphs [D]. Xi'an Shiyou University, 2022.]

[28] 赵继贵, 钱育蓉, 王魁, 等. 中文命名实体识别研究综述[J]. 计算机工程与应用, 2024, 60(01): 15−27.[ZHAO J G, QIAN Y R, WANG K, et al. A review of Chinese named entity recognition research [J]. Computer Engineering and Applications, 2024, 60(01): 15−27.]

[29] 刘炳旭. 油气勘探领域命名实体识别的研究与实现[D]. 中国石油大学(北京), 2021.[LIU B X. Research and Implementation of Named Entity Recognition in the Field of Oil and Gas Exploration [D]. China University of Petroleum (Beijing), 2021.]

[30] 程志刚. 基于规则和条件随机场的中文命名实体识别方法研究[D]. 华中师范大学, 2015.[CHENG Z G. Research on Chinese named entity recognition method based on rules and conditional random fields [D]. Central China Normal University, 2015.]

[31] YY B, WEI J. IDCNN-CRF-based domain named entity recognition method[C]//2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT. IEEE, 2020: 542−546.

[32] 刘文聪, 张春菊, 汪陈, 等. 基于BiLSTM-CRF的中文地质时间信息抽取 [J]. 地球科学进展, 2021, 36 (02): 211−220.[LIU W C, ZHANG C J, WANG C, et al. Chinese geological time information extraction based on BiLSTM-CRF [J]. Progress in Earth Science, 2021, 36 (02): 211−220.]

(编辑 马桂霞)