

基于RoBERTa-wwm的飞行超限事件 命名实体识别

冯军红¹, 余雪², 廖颖馨¹, 丁昱婷²

(1. 中国民用航空飞行学院民航监察员培训学院; 2. 中国民用航空飞行学院飞行技术学院, 四川 广汉 618307)

摘要: 针对传统模型因领域适应性差, 无法有效捕捉民航专业语义、界定实体边界, 导致识别精度不足的问题。将2 315份飞行超限事件调查报告作为数据源, 以构建飞行超限事件数据集, 并提出一种RoBERTa-wwm-BiLSTM-Att-CRF命名实体识别模型。首先, 基于RoBERTa-wwm在大规模语料上预训练获得的深层语言表征, 生成高质量的字级上下文嵌入向量; 其次, 结合双向长短期记忆网络(BiLSTM)捕捉上下文序列特征, 并通过注意力机制(Att)聚焦关键信息; 最后, 通过条件随机场(CRF)解码全局最优标注序列。在自建数据集上的实验表明, 所提模型的F1值达到91.89%, 性能显著优于现有方法, 能有效解决数据集中的实体边界模糊等问题, 提升命名实体识别效果, 为自动构建飞行超限事件知识图谱提供了方法支撑。

关键词: 飞行超限事件; 命名实体识别; RoBERTa-wwm; 注意力机制; 条件随机场

DOI: 10.11907/rjtk.251720

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(XXXX)0XX-0001-08



扫描二维码阅读全文:

Named Entity Recognition for Flight Exceedance Events Based on RoBERTa-wwm

FENG Junhong¹, YU Xue², LIAO Yingxin¹, DING Yuting²

(1. Civil Aviation Inspector Training Academy, Civil Aviation Flight University of China;
2. Flight Technology College, Civil Aviation Flight University of China, Guanghan 618307, China)

Abstract: Due to poor domain adaptability, traditional models are unable to effectively capture civil aviation professional semantics and define entity boundaries, resulting in insufficient recognition accuracy. Using 2 315 investigation reports on flight over limit events as the data source, a dataset of flight over limit events was constructed, and a RoBERTa-wwm-BiLSTM-Att-CRF named entity recognition model was proposed. Firstly, based on RoBERTa-wwm pre trained deep language representations on large-scale corpora, high-quality word level context embedding vectors are generated; Secondly, combining bidirectional long short-term memory network (BiLSTM) to capture contextual sequence features, and focusing on key information through attention mechanism (Att); Finally, the globally optimal annotation sequence is decoded using Conditional Random Fields (CRF). Experiments on self built datasets have shown that the F1 value of the proposed model reaches 91.89%, which is significantly better than existing methods. It can effectively solve problems such as fuzzy entity boundaries in the dataset, improve named entity recognition performance, and provide methodological support for automatically constructing a knowledge graph of flight over limit events.

Key Words: flight exceedance events; named entity recognition; RoBERTa-wwm; attention mechanism; conditional random field

0 引言

民航飞行超限事件调查报告, 为航空企业和监管部门

识别潜在安全隐患提供了重要的决策依据, 但调查报告多为非结构化文本, 信息分布零散且表述不一, 导致传统人工分析方法难以高效提取有效信息。

知识图谱技术具备结构化表达与关联分析能力, 因此

收稿日期: 2025-10-23

作者简介: 冯军红(1981-), 男, 硕士, 中国民用航空飞行学院民航监察员培训学院副教授, 研究方向为民航航行新技术; 余雪(2001-), 女, 中国民用航空飞行学院飞行技术学院硕士研究生, 研究方向为知识图谱; 廖颖馨(1988-), 女, 中国民用航空飞行学院民航监察员培训学院硕士研究生, 研究方向为空管安全管理; 丁昱婷(2002-), 女, 中国民用航空飞行学院飞行技术学院硕士研究生, 研究方向为飞行员绩效评估。

能提炼并整合调查报告中分散的关键信息^[1,2]。命名实体识别(Named Entity Recognition, NER)作为知识图谱构建的核心前置环节,承担着在非结构化文本中,精准抽取语义实体和结构化存储的重要任务^[3]。然而,飞行超限事件调查报告的文本特性(实体边界模糊、专业术语繁多等)对NER技术的识别精度提出了更高要求。

早期,NER方法基于人工设计的规则与模板,但过度依赖专家经验,因此领域迁移能力弱、泛化性能差^[4]。随后,统计机器学习模型逐渐成为主流,例如隐马尔可夫模型、条件随机场、支持向量机^[5-7]。这些方法将NER建模为序列标注问题,在通用领域文本中取得了较好效果,但严重依赖人工特征工程,在面对结构复杂、术语多样的领域文本时,识别性能受限。Han等^[8]提出结合多层CRF解码与分割感知的关系式图卷积网络,通过“分割边界信息+二次解码”减少误差传递并建模类别间的依赖关系,以稳步提升模型性能。Li等^[9]提出一种基于全局指针与对抗训练机制的中文命名实体识别模型,在解码阶段引入全局边界建模与对抗扰动优化,有效提升了模型对实体边界的敏感性与整体鲁棒性,在中文NER任务上的性能优秀。Zhang等^[10]提出一种融合全局指针与对抗训练的模型,以进一步增强模型在复杂句式与模糊边界下的实体识别能力,在多项中文数据集的识别精度较高。Yazdani等^[11]提出Fourier网络联合框架,联合低复杂度的频域token-mixing与选择性池化抽取实体与关系,在临床ADE基准上训练更快、显存更省,为高效进行NER提供了可行路径。

近年来,预训练语言模型的兴起推动了NER技术的发展。He等^[12]提出一种基于Prompt-BERT的命名实体识别方法,在词级别中注入提示信息来增强上下文关联性,有效提升了模型对实体边界与语义一致性的捕捉能力。王月等^[13]针对警情领域实体难以识别的问题,构建BERT-BiLSTM-Attention-CRF模型,在相应任务中F1值达到91%。许力等^[14]在生物医学领域构建了BERT-BiLSTM-CRF模型。方睿等^[15]提出一种融合BERT的事件抽取方法KG-PMEE。上述方法都有效提升了模型性能,但在中文NER任务中,由于汉语缺乏显式的词界标记,BERT系列模型往往难以充分捕捉词汇级别的完整语义,对专业术语的建模能力不足。RoBERTa-wwm模型在预训练阶段引入全词掩码策略,增强了模型对词汇语义的学习能力,在多项中文NER任务中的性能更强^[16,17]。

尽管,预训练模型已取得了显著成果,但民航领域文本存在实体边界模糊、句式结构复杂等问题,现有模型的性能仍然不足以满足需求。为此,王红等^[18]针对民航突发事件文本,提出基于BiLSTM-CRF的识别模型,并融合领域词典辅助标注,以提升对复合型实体的识别效果。许雅玺等^[19]结合BERT词向量与航空不安全领域词典嵌入,增强了模型对实体边界的感知能力。然而,现有方法大多依赖外部词典,未充分挖掘预训练模型内部的深层语义信息

与长距离依赖关系,因此模型在语义表征和标签序列推断方面仍面临挑战。

为此,本文提出一种RoBERTa-wwm-BiLSTM-Att-CRF命名实体识别模型。首先,充分分析飞行超限事件调查报告的数据特点,定义命名实体类别以构建飞行超限事件数据集,并利用开源标注工具对数据集进行标注;其次,针对传统模型在实体边界信息捕获方面的不足,基于RoBERTa-wwm预训练语言模型,提出融合双向长短期记忆网络(Bidirectional Long Short Term Memory Network, BiLSTM)、注意力机制与条件随机场(Conditional Random Field, CRF)的命名实体识别方法,旨在精确识别数据集内的命名实体,解决数据集的非结构化文本问题。

1 数据集与预处理

1.1 数据采集与清洗

研究选取2017年1月—2024年10月,中国民用航空发生的2315份飞行超限事件调查报告为研究对象,包含时间、机型、航段、原因分析、调查结论等。通过对原始数据进行预处理,去除冗余信息和无细节后,构建了本文研究的飞行超限事件数据集。数据清洗过程为:

(1)去标识化处理。基于民航数据隐私保护要求,采用正则表达式匹配与人工核验相结合的方式,对报告中涉及的航空公司、航空器注册号和航班号等敏感信息进行统一掩码处理。

(2)关键信息提取。针对超限事件调查报告特有的章节标题结构特征,采用文本抽取算法自动保留事件经过、调查结论等核心内容,剔除非相关内容。

(3)文本规范化。标准化时间表达式,例如将2023-05-01统一为2023年5月1日。

1.2 半自动化标注

为了提升命名实体识别任务中非结构化文本数据的标注效率与准确性,采用了“领域词典引导+人工校验”的两阶段半自动标注策略,如图1所示。

首先,收集部分表格数据,并从中提取所需实体词条,以构建一个包含多个实体类别的领域词典,每个实体类别

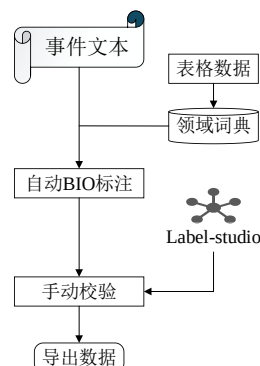


Fig. 1 Annotation process

图1 标注流程

对应一个子词表;其次,将领域词典应用于非结构化文本的初步自动标注任务。具体而言,基于正则表达式匹配与字符串搜索等技术,对原始文本进行词典实体定位与识别,并依据实体类别为词项标注对应的 BIO 标签。当自动标注完成后,将生成的预标注文件格式化为 JSON 格式,以便导入标注平台。

在人工标注阶段,借助 Label Studio 平台对自动标注的结果进行人工校验与修正。Label Studio 是一款开源的文本标注工具,能有效支持自然语言处理任务的文本标注工作。本文标注的实体类型包括:航空公司、起飞机场、目的机场、事件地点、事件原因、事发阶段、事件结果、事件时间及飞机型号等9类。在标注完成后,将结果导出为 JSON 格式文件,以详细记录每个标注实体的位置索引信息,如图 2 所示(不同实体类型通过不同颜色区分)。

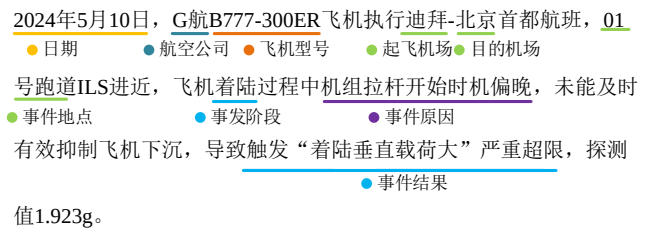


Fig. 2 Entity annotation example
图2 实体标注示例

1.3 标注数据转换

本文利用 Python 编程语言对导出的 JSON 文件数据进行处理。同时,采用 BIO 实体标注法进行数据转换,旨在为后续命名实体识别模型训练构建标准化的数据集。其中,B表示实体的开始位置,I表示实体内部,O表示非实体词汇。实体标注数量的具体统计结果如表 1 所示。

Table 1 Entity quantity statistics
表1 实体数量统计

实体类型	数量
航空公司	2 320
起飞机场	2 340
目的机场	2 320
事件地点	1 677
事件原因	2 569
事发阶段	1 973
事件结果	3 292
飞机型号	2 334
事件时间	2 319
总量	21 144

2 模型设计

2.1 模型框架

RoBERTa-wwm-BiLSTM-Att-CRF 模型采用层级式架构,依次实现文本深度语义编码、上下文序列特征提取、关键语义信息增强与全局最优标签序列预测,如图 3 所示。

由此可见,模型整体架构包含 4 个核心组件:①RoBERTa-wwm 编码层基于全词掩码策略的预训练模型,用于生成融合上下文信息的动态词表征;②BiLSTM 序列建模层通过双向循环网络捕获文本的局部上下文依赖特征;③Attention 注意力机制层动态加权关键语义信息,突出实体关键信息;④CRF 解码层建模标签转移约束,以预测全局最优的标签序列。模型 4 层架构形成了从原始文本输入到实体标签输出的端到端映射流程,通过多组件协同机制,可有效应对飞行超限事件数据集中的实体边界模糊、术语密集等识别难点。

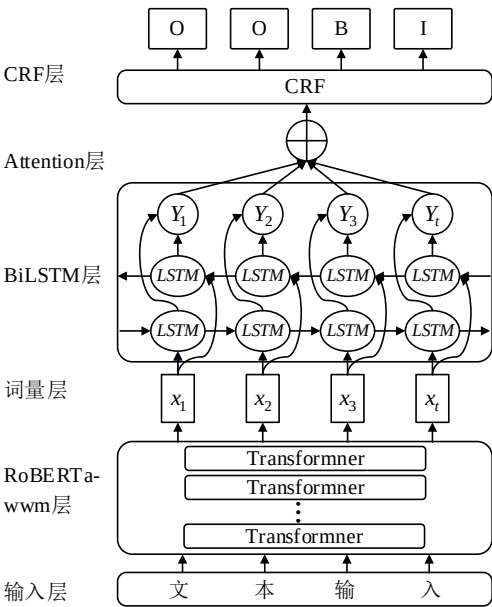


Fig. 3 RoBERTa-wwm-BiLSTM-Attention-CRF model
图3 RoBERTa-wwm-BiLSTM-Attention-CRF 模型

2.2 RoBERTa-wwm 预训练模型层

RoBERTa-wwm 由哈工大讯飞联合实验室发布,旨在为中文自然语言处理任务提供更好的性能。该模型基于 BERT 架构,继承了 BERT 的双向 Transformer 结构,以多头自注意力 (Multi-Head Self-Attention) 机制为核心,通过计算句子中任意两个词之间的关联性来获取句中丰富的语义信息^[20]。

RoBERTa-wwm 模型的工作原理为:利用 Transformer 结构构建多层双向编码器网络,以读取整个文本序列。该模型每层都可集成上下文信息,如图 4 所示。其中, $T = \{t_1, t_2, t_3, \dots, t_n\}$ 为输入文本; $H = \{h_1, h_2, h_3, \dots, h_n\}$ 为模型输出目标单词的词向量,含有 RoBERTa-wwm 在预训练阶段获得的先验语义知识。

RoBERTa-wwm 相较于传统 BERT 模型,改进了掩蔽语言模型 (Masked Language Model, MLM) 策略,采用全词遮掩机制,在预训练阶段先对句子进行分词,而非仅掩蔽单个字 (字级别 Mask)。该改进使模型在学习中能更好地捕捉词汇的整体语义信息,解决了 BERT 仅能获取字级别语义的局限性,提升了对中文复合词、专业术语的识别精度。

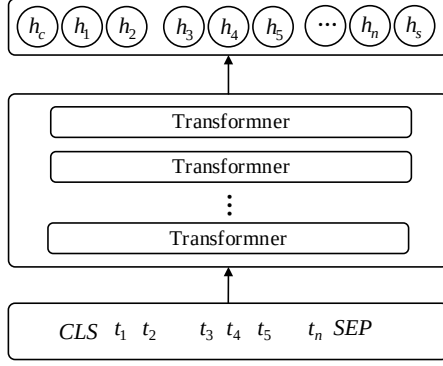


Fig. 4 RoBERTa-wwm model architecture

图4 RoBERTa-wwm模型结构

同时, RoBERTa-wwm模型进一步优化了训练流程, 移除了下一句预测(Next Sentence Prediction, NSP)任务, 增加了训练时的序列长度, 使模型能处理更长的文本输入。此外, 模型使用动态掩码机制, 即每次训练时掩蔽的词汇都会发生变化, 使模型能在多次迭代中逐渐适应不同的掩蔽策略, 以更精确地捕捉语义多样性和上下文信息。

本文研究中, RoBERTa-wwm层主要用于对输入的飞行超限事件文本进行语义编码, 以生成每个词的深度上下文语义表示。

2.3 BiLSTM序列建模层

为了更好地捕获文本序列中的局部依赖关系, 在RoBERT-wwm编码层输出的基础上引入BiLSTM。BiLSTM网络由两个方向相反的长短期记忆(Long Short-Term Memory, LSTM)模型组成(见图5)。LSTM是一种改进型循环神经网络(Recurrent Neural Network, RNN), 旨在解决传统RNN处理长序列数据时易出现的梯度消失与梯度爆炸问题^[21]。具体做法为: LSTM通过引入遗忘门、输入门与输出门和细胞状态, 选择性记忆和遗忘信息, 对信息流动进行精确控制, 以增强模型捕捉长期依赖关系的能力^[22]。

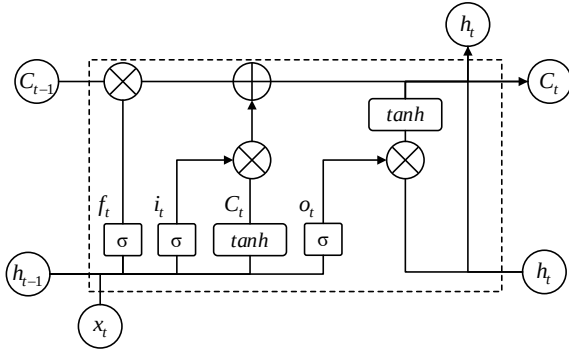


Fig. 5 LSTM model architecture

图5 LSTM模型结构

遗忘门根据前一时刻的隐藏层状态 h_{t-1} 和当前时刻的输入序列 x_t , 计算一个0~1之间的权重向量, 以决定先前细胞状态中哪些信息需要保留或丢弃, 0表示完全删除, 1表示完全保留。计算如公式(1)所示。

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \quad (1)$$

式中: f_t 表示当前时刻 t 的遗忘门输出; W_f, b_f 分别表示遗忘门的权重矩阵和偏置项; σ 代表tanh激活函数。

输入门负责筛选当前时刻的新信息, 决定需要保存在神经元中的信息。首先, 通过tanh函数生成一个0~1之间的门控值 i_t , 表示当前输入信息的重要性, i_t 接近1则允许大量新信息进入细胞状态, i_t 接近0则抑制无关输入; 其次, 通过tanh层生成时刻 t 的候选记忆单元状态 $\tilde{C}_t$; 最后, 结合遗忘门的输出对细胞状态进行更新, 得到新的细胞状态 C_t 。具体运算过程如式(2)~式(4)所示。

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_c \times [h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (4)$$

式中: W_i, b_i 分别为输入门的权重和偏置项; W_c, b_c 为生成候选记忆单元的权重矩阵和偏置; tanh为双曲正切激活函数; $\odot$ 表示元素逐位相乘。

输出门决定当前时刻 t 的隐藏状态 h_t , 即从细胞状态 C_t 中提取哪些信息输出。首先, 通过 σ 决定哪些神经元状态需要被输出; 其次, 将这些神经元状态输入tanh进行处理, 再与 o_t 相乘, 生成最终的输出值 h_t , 该值也作为下一个时间步的隐藏层输入。计算过程如式(5)、式(6)所示。

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t \odot \tanh(C_t) \quad (6)$$

式中: o_t 为输出门在时刻 t 的输出状态; W_o, b_o 分别为输出门的权重矩阵和偏置项。

值得注意的是, 传统LSTM采用单向信息流处理模式, 仅考虑从序列起始到结束的正向时序特征, 因此可能导致部分时序信息未被充分利用。BiLSTM创新性地整合了两个独立的LSTM层: 一层按时间正序处理输入序列; 另一层按时间逆序处理同一序列。正向LSTM主要捕捉从过去到当前的依赖特征, 反向LSTM则捕捉从未来到当前的依赖特征。图6为BiLSTM的模型结构。

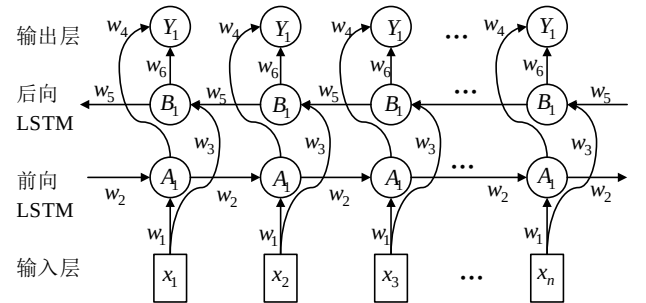


Fig. 6 BiLSTM model architecture

图6 BiLSTM模型结构

BiLSTM的输出过程如式(7)~式(9)所示。通过拼接或组合两个方向的隐藏状态, 以获得同时包含前向与后向信息的序列表示。这种双向序列建模方式, 可有效增强模型对短距离特征及上下文信息的捕获能力。

$$A_i = f_1(w_1 x_i + w_2 A_{i-1}) \quad (7)$$

$$B_i = f_2(w_3 x_i + w_5 B_{i+1}) \quad (8)$$

$$Y_i = f_3(w_4 A_i + w_6 B_i) \quad (9)$$

式中: A_i 为正向 LSTM 在时间步 i 的隐藏状态; B_i 为反向 LSTM 在时间步 i 的隐藏状态; Y_i 为 BiLSTM 在时间步 i 的输出; f_1, f_2, f_3 分别为不同层之间的激活函数; w_k 表示对应连接的权重参数。

2.4 Attention 注意力机制层

尽管 BiLSTM 能捕获序列特征, 但对长序列中的关键实体信息的捕获能力有限。为此, 在 BiLSTM 层引入注意力机制, 通过学习注意力权重矩阵, 赋予文本序列中不同词汇不同的重要性, 使模型在实体识别时更聚焦事件中的关键词汇和语义信息^[23]。

首先, 给定解码器在当前时刻的隐状态向量 q 、编码器输出的隐层表示序列 $\{s_1, s_2, \dots, s_n\}$, 通过打分函数计算查询向量与输入表示之间的相关性得分。

$$a_i = \text{score}(q, s_i) \quad (10)$$

其次, 将所有相关性得分 $\{a_1, a_2, \dots, a_n\}$ 输入 Softmax 函数, 得到标准化的注意力权重。

$$\alpha_i = \frac{\exp(a_i)}{\sum_{j=1}^n \exp(a_j)}, \alpha_i \geq 0, \sum_{i=1}^n \alpha_i = 1 \quad (11)$$

最后, 根据注意力分布对输入隐层表示进行加权求和, 得到上下文向量。该向量能动态捕捉与当前预测最相关的实体信息, 并与解码器隐态共同作用于后续的输出过程。注意力机制的结构图 7 所示。

$$c = \sum_{i=1}^n \alpha_i s_i \quad (12)$$

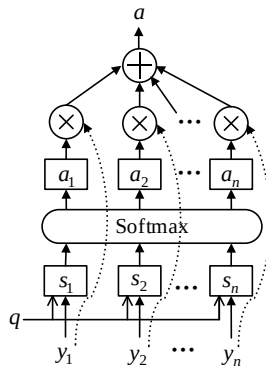


Fig. 7 Structure of attention mechanism

图 7 注意力机制的结构

在命名实体识别任务中, Attention 层以双向 LSTM 的隐藏状态为输入, 通过计算当前位置与序列中各位置之间的相关性得分, 以获得一组归一化的注意力权重, 并据此对全局上下文进行加权求和来生成上下文向量。该机制能在建模上下文信息时, 有效突出与目标实体最相关的关键词, 缓解长序列中关键信息被稀释的问题。

2.5 CRF 解码层

实体识别任务本质上是一种序列标注问题, 标签之间存在明显的依赖关系, 例如“B-飞机型号”通常紧跟“I-飞

机型号”, 而非其他标签。CRF 通过考虑标签之间的上下文约束, 实现标签序列的全局最优解码^[24,25]。具体地, CRF 层根据前一层的输出特征序列, 利用序列标签转移概率模型优化整个序列的标签选择, 如式 (13) 所示。

$$S(X, y) = \sum_{i=1}^n W_{y_{i-1}, y_i} + \sum_{i=1}^n P_{i, y_i} \quad (13)$$

式中: $S(X, y)$ 表示输入序列 X 与标签序列 y 的得分; W_{y_{i-1}, y_i} 表示标签转移概率; P_{i, y_i} 表示当前词对应的发射概率。

式 (13) 由两个关键部分组成: ① 发射得分项 $\sum_{i=1}^n P_{i, y_i}$ 量化输入序列中每个词汇单元与其对应标签的匹配程度, 数值来源于前层神经网络输出的未归一化对数概率; ② 转移得分项 $\sum_{i=1}^n W_{y_{i-1}, y_i}$ 通过可学习的标签转移矩阵显式编码标签序列的语法规则, 确保输出标签序列符合预定义的约束条件。

在模型训练阶段, 式 (13) 通过 softmax 归一化为条件概率分布, 采用最大似然估计进行优化; 在模型解码阶段, 通过动态规划的维特比算法, 高效搜索得分最高的合法标签序列^[26]。这种设计使 CRF 能有效平衡局部预测准确性和全局序列合理性, 提升命名实体识别等序列标注任务的性能。

3 实验结果与分析

3.1 数据集与评价指标

将包含 2 315 条飞行超限事件的数据集, 按 8:1:1 的比例划分为训练集、验证集和测试集, 用于模型训练与评估。为了全面评估模型性能, 采用精确率 (Precision, P)、召回率 (Recall, R)、F1 值 (F1-score) 作为主要评价指标, 计算公式如式 (14)~式 (16) 所示。其中, 精确率用于衡量模型预测结果中真正实体所占的比例; 召回率衡量所有实际实体中被模型成功预测出的比例; F1 值则是精确率与召回率的调和平均数, 以综合评估模型整体性能。

$$P = \frac{TP}{TP + FP} \times 100\% \quad (14)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (15)$$

$$F1 = \frac{2 \times P \times R}{(P + R)} \times 100\% \quad (16)$$

式中: TP (True Positive) 表示正确识别的实体数; FP (False Positive) 表示被错误识别为实体的非实体数; FN (False Negative) 表示被漏识别的实体数。

3.2 实验环境与模型参数

实验硬件环境为: NVIDIA Quadro P2200 显卡 (5 GB 显存)、Intel Xeon W-2104 处理器 (3.20 GHz, 4 核) 与 32 GB DDR4 内存, 操作系统为 Windows 10 专业版 64 位。软件环

境为:Python 3.7.4与PyTorch 1.3.1,主要依赖库为PyTorch-transformers 1.2.0、PyTorch-crf 0.7.2与tensorboardX等。实验环境如表2所示。

Table 2 Experimental environment

表2 实验环境

实验环境	环境配置
操作系统	Windows 10
CPU	Intel Xeon W-2104 CPU @ 3.20GHz
内存/GB	32
GPU	NVIDIA Quadro P2200
Python 版本	3.7.4
PyTorch 版本	1.3.1

本文模型的预训练语言模型为“roberta-wwm-ext”中文模型。考虑到5 GB显存的资源约束,将最大输入序列长度设为128、训练批大小设为2,并配合梯度累积以获得等效更大的batch;学习率设置为 $3e-5$,训练轮数为32,BiLSTM隐藏维度为128,其余参数保持默认设置。具体参数如表3所示。

Table 3 Model parameters

表3 模型参数

参数	对应值
最大序列长度	128
batch size	8
learning rate	$2e-5$
epoch	32
BiLSTM隐藏层维度	128
Dropout 概率	0.1 (attention+hidden)
优化器	AdamW

3.3 超参数选择与训练动态

为了在不同超参数配置下获得最优性能,采用结合早停策略和最佳轮次选择的方法,以自动确定每个注意力头数对应的最优训练轮次^[27]。首先,在不同注意力头数下分别进行模型训练,每个模型都训练至最大32轮;其次,每一轮训练结束后在验证集上计算模型的F1值,以反映该阶段模型的泛化性能;最后,根据验证集表现选取不同注意力头数下,F1值最高的一轮作为该配置的最佳训练轮次,并将所有头数配置下的最优验证集F1值进行比较,选择最好的组合以确定全局最优注意力头数与训练轮次的配对方案。该方法能在避免过拟合的同时,更高效地找到模型性能峰值点。

实验发现,在注意力头数为4,训练轮次为21时,模型效果最优。图8为注意力头数为4时,不同训练轮次下的模型性能变化情况。图9为训练轮次为21时,不同注意力头数下的模型性能变化情况。

3.4 实验结果

表4为RoBERTa-wwm-BiLSTM-Att-CRF模型在飞行超限事件数据集中的命名实体识别结果。由此可知,模型在“事件时间”实体识别上表现最优,准确率与召回率均达到100%,表明该类别实体在文本中的表述形式高度统一。“航空公司”“目的机场”“起飞机场”“飞机型号”实体的F1

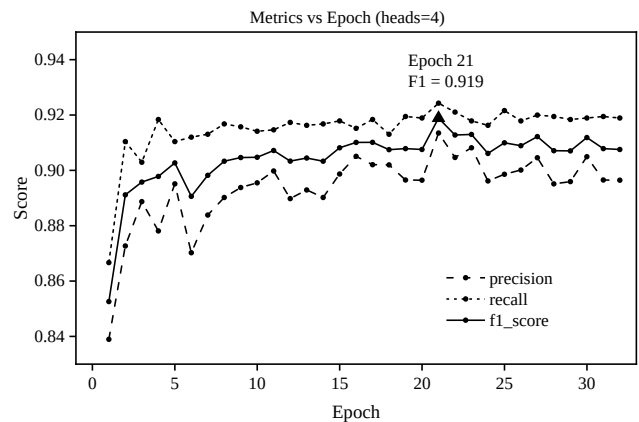


Fig. 8 Model performance changes under different training epochs (with 4 attention heads)

图8 不同训练轮次下的模型性能变化(注意力头数为4)

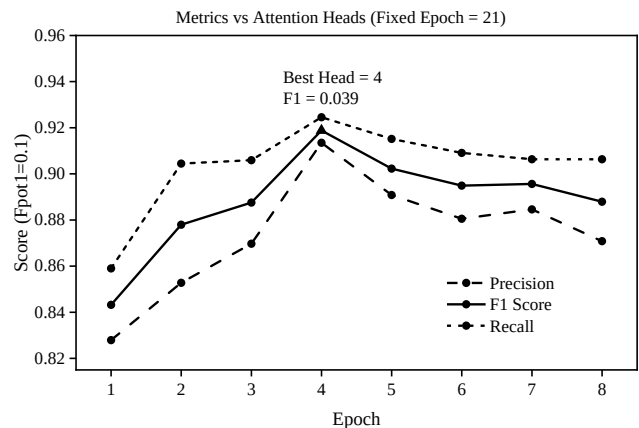


Fig. 9 Performance changes of models under different numbers of attention heads (training epochs of 21)

图9 不同注意力头数下的模型性能变化(训练轮次为21)

值均超过95%,模型在识别这类实体时具有良好的泛化能力。相较而言,“事件地点”“事件原因”“事发阶段”的F1值相对较低,主要原因在于这些实体在自然语言表述中常存在歧义或句法结构复杂多变,对模型识别性能产生了影响。

Table 4 Named entity recognition results

表4 命名实体识别结果

(%)

实体类型	准确率	召回率	F1值
事件原因	80.00	80.45	80.22
事件地点	76.79	79.14	77.95
事件时间	100.00	100.00	100.00
事件结果	88.44	87.93	88.18
事发阶段	81.58	84.24	82.89
目的机场	95.00	96.61	95.80
航空公司	99.57	100.00	99.78
起飞机场	95.47	97.89	96.67
飞机型号	95.78	95.78	95.78
整体指标	91.35	92.43	91.89

3.5 比较实验

表5为BiLSTM、BERT、GRU-CRF等模型在飞行超限事件命名实体识别任务上的性能,BERT采用的是“Chi-

nese-base-Bert”中文预训练模型。由此可知, 基于 Transformer 架构的预训练模型显著优于其他传统深度学习模型。其中, BiLSTM 作为常用的命名实体识别模型, 在飞行超限事件数据集上的 F1 值为 75.82%, 尽管在召回率方面表现尚可, 但精确率仅为 73.84%, 表明该模型在实体边界识别和类别判定上存在不足, 主要原因为: BiLSTM 模型采用的静态词嵌入机制, 难以根据上下文语境动态调整词义, 导致词汇歧义必须完全依赖后续编码层进行解析, 既增加了模型学习负担, 又限制了模型捕捉深层语义和长程依赖关系的能力。相较而言, BERT 与 RoBERTa-wwm 凭借 Transformer 自注意力机制架构和掩码语言模型的预训练, 能生成深度上下文相关的动态词表示, 有效解决了词汇歧义问题, 可为实体识别提供更精准的语义。

Table 5 Experimental results of different models			
表 5 不同模型的实验结果 (%)			
模型	Precision/%	Recall/%	F1/%
CNN-CRF	67.26	75.66	71.22
GRU-CRF	75.89	70.95	73.34
BiLSTM	73.84	77.91	75.82
BiLSTM-CRF	77.78	83.7	80.63
BERT-CRF	87.28	86.78	87.03
BERT-BiLSTM-CRF	89.87	90.4	90.14
RoBERT-wwm-BiLSTM-Att-CRF	91.35	92.43	91.89

在结构相同的下游任务中, 以 RoBERTa-wwm 为基座的模型均优于对应 BERT 基座的模型。由表 5 可知, BERT-CRF 的 F1 值为 87.03%, RoBERTa-CRF 的 F1 值为 88.18%, 相较于 BERT-CRF 提升 1.15%, RoBERTa-wwm-BiLSTM-CRF 模型的识别效果优于 BERT-BiLSTM-CRF, 主要原因为: RoBERTa-wwm 在预训练阶段通过移除 NSP 任务、采用全词掩码策略、扩充语料规模及引入动态掩码技术等, 显著增强了模型的中文语义表征能力。

3.6 消融实验

为了验证各模块的有效性, 通过消融实验确定各组件对模型性能的贡献程度, 如表 6、图 10 所示。由此可知, 仅使用 RoBERT-wwm-Att-CRF 时, 模型的 F1 值仅为 80.22%, 说明 BiLSTM 在捕捉序列特征方面发挥了关键作用; 去除 CRF 模块后, 模型性能下降明显, 说明 CRF 在命名实体识别任务中具有重要作用; 去除 Attention 后, 性能下降幅度相对较小。整体而言, BiLSTM、CRF 是提升模型性能的核心组件, 而 Attention 主要起优化作用。

Table 6 Results of ablation experiments			
表 6 消融实验结果 (%)			
模型	Precision	Recall	F1
RoBERT-wwm-Att-CRF	80.00	80.45	80.22
RoBERT-wwm-BiLSTM-Att	88.44	87.93	88.18
RoBERT-wwm-BiLSTM-CRF	90.43	91.25	90.84
RoBERT-wwm-BiLSTM-Att-CRF	91.35	92.43	91.89

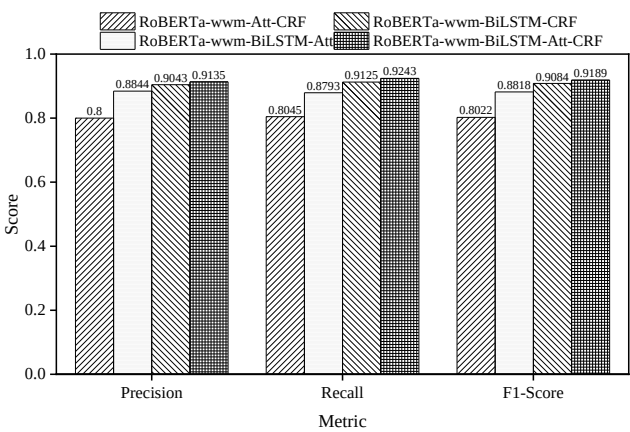


Fig. 10 Visualized results of ablation studies

图 10 消融实验可视化结果

综上, RoBERTa-wwm 用于提供高质量的深层语义表示; CRF 层通过建模标签转移约束, 以确保输出序列的全局合理性; BiLSTM 增强了模型捕捉上下文序列中局部依赖关系的能力; Attention 机制进一步提升了模型对关键特征的聚焦能力。四者有效融合, 可使 RoBERT-wwm-BiLSTM-Att-CRF 在飞行超限事件数据集上的精确率、召回率与 F1 值均达到最优水平, 验证了本文模型架构的有效性与合理性。

4 结语

本文针对飞行超限事件命名实体识别任务中存在的实体边界模糊与专业术语繁多等问题, 提出一种 RoBERTa-wwm-BiLSTM-Att-CRF 命名实体识别模型。首先, 以 2 315 份飞行超限事件调查报告为数据源, 构建了包含事发阶段、事件原因、事件结果等 9 类核心实体的民航专业领域数据集, 解决了该研究场景下数据匮乏的难题; 其次, 形成语义增强—特征聚焦—全局约束的创新架构, 在中文语义理解、上下文特征捕获、实体标签预测等方面表现出了良好的性能。

在自建数据集上的实验表明, 所提模型在实体识别任务中的精度相较于现有模型性能最优。消融实验表明, 所提模型中的 4 个部分对提升模型精确率、召回率与 F1 值均具有积极作用, 为自动构建飞行超限事件知识图谱奠定了重要基础。

参考文献:

[1] JI S, PAN S, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494–514.

[2] ZHONG L, WU J, LI Q, et al. A comprehensive survey on automatic knowledge graph construction[J]. ACM Computing Surveys, 2023, 56(4): 1–62.

[3] HU Z, HOU W, LIU X. Deep learning for named entity recognition: a survey[J]. Neural Computing and Applications, 2024, 36(16): 8995–9022.

- [4] ZHAO J G, QIAN Y R, WANG K, et al. Survey of Chinese named entity recognition research [J]. *Journal of Computer Engineering & Applications*, 2024, 60(1): 15–27.
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(1): 15–27.
- [5] MOR B, GARHWAL S, KUMAR A. A systematic review of hidden Markov models and their applications [J]. *Archives of Computational Methods in Engineering*, 2021, 28(3): 1429–1448.
- [6] LI J, SUN A, HAN J, et al. A survey on deep learning for named entity recognition [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2020, 34(1): 50–70.
- [7] IMAM A T, ALHROOB A, ALZYADAT W. SVM machine learning classifier to automate the extraction of SRS elements [EB/OL]. https://thesai.org/Downloads/Volume12No3/Paper_22-SVM_Machine_Learning_Classifier.pdf.
- [8] HAN D, WANG Z, LI Y, et al. Segmentation-aware relational graph convolutional network with multi-layer CRF for nested named entity recognition [J]. *Complex & Intelligent Systems*, 2024, 10(6): 7893–7905.
- [9] LI H, CHENG M, YANG Z, et al. Named entity recognition for Chinese based on global pointer and adversarial training [J]. *Scientific Reports*, 2023, 13(1): 3242.
- [10] ZHANG M, LI Y, ZHANG Z, et al. Named entity recognition for Chinese scientific literature based on adversarial training and dual-masked global pointer [J]. *Complex & Intelligent Systems*, 2025, 11(10): 1–18.
- [11] YAZDANI A, PROIOS D, ROUHIZADEH H, et al. Efficient joint learning for clinical named entity recognition and relation extraction using Fourier networks: a use case in adverse drug events [DB/OL]. <https://arxiv.org/abs/2302.04185>.
- [12] HE Q, CHEN G, SONG W, et al. Prompt-based word-level information injection BERT for Chinese named entity recognition [J]. *Applied Sciences*, 2023, 13(5): 3331.
- [13] WANG Y, WANG M X, ZHANG S, et al. Alarm text named entity recognition based on BERT [J]. *Journal of Computer Applications*, 2020, 40(2): 535–540.
王月, 王孟轩, 张胜, 等. 基于 BERT 的警情文本命名实体识别[J]. *计算机应用*, 2020, 40(2): 535–540.
- [14] XU L, LI J H. Biomedical named entity recognition based on BERT and BiLSTM-CRF [J]. *Computer Engineering & Science*, 2021, 43(10): 1873–1879.
许力, 李建华. 基于 BERT 和 BiLSTM-CRF 的生物医学命名实体识别[J]. *计算机工程与科学*, 2021, 43(10): 1873–1879.
- [15] FANG R, CUI L Z. An event extraction method in specific domains combining knowledge graph [J]. *Software Guide*, 2025, 24(9): 62–69.
方睿, 崔良中. 融合知识图谱的领域事件抽取方法[J]. *软件导刊*, 2025, 24(9): 62–69.
- [16] GAO F, ZHANG L, WANG W, et al. Named entity recognition for equipment fault diagnosis based on RoBERTa-wwm-ext and deep learning integration [J]. *Electronics*, 2024, 13(19): 3935.
- [17] ZHANG H, SHAFIQ M O. Survey of transformers and towards ensemble learning using transformers for natural language processing [J]. *Journal of Big Data*, 2024, 11(1): 25.
- [18] WANG H, LI H F, DI S. Entity recognition method of civil aviation emergency [J]. *Computer Applications and Software*. 2020, 37(3): 166–172.
王红, 李浩飞, 邸帅. 民航突发事件实体识别方法研究[J]. *计算机应用与软件*, 2020, 37(3): 166–172.
- [19] XU Y X, MENG T Y, WANG X, et al. Named entity recognition of aviation unsafe events embedded with fusion domain dictionary [J]. *Science Technology and Engineering*, 2024, 24(8): 3284–3290.
许雅玺, 孟天宇, 王欣, 等. 融合领域词典嵌入的航空不安全事件命名实体识别[J]. *Science Technology & Engineering*, 2024, 24(8): 3284–3290.
- [20] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for chinese BERT [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, 29: 3504–3514.
- [21] WANG Z, SU X, DING Z. Long-term traffic prediction based on LSTM encoder-decoder architecture [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2020, 22(10): 6561–6571.
- [22] BECK M, PöPPEL K, SPANRING M, et al. XLSTM: extended long short-term memory [J]. *Advances in Neural Information Processing Systems*, 2024, 37: 107547–107603.
- [23] PAN Y, YANG J, MING H, et al. Few-Shot named entity recognition via label-attention mechanism [C]// 9th International Conference on Computing and Artificial Intelligence, 2023: 466–471.
- [24] MA M, YANG Q, ESKAR A, et al. Chinese named entity classification based on word vectors and conditional random fields [J]. *Computer Engineering and Design*, 2020, 41(9): 2515–2522.
- [25] GUO Y, LIU H C, LIU F J, et al. Chinese named entity recognition with multi-network fusion of multi-scale lexical information [J]. *Journal of Electronic Science and Technology*, 2024, 22(4): 100287.
- [26] SÄRKKÄ S, GARCÍA-FERNÁNDEZ Á F. On the temporal parallelisation of the viterbi algorithm [C]// 31st European Signal Processing Conference, 2023: 2018–2022.
- [27] LI J, WANG X, TU Z, et al. On the diversity of multi-head attention [J]. *Neurocomputing*, 2021, 454: 14–24.

(责任编辑: 刘嘉文)