文章编号:1671-4598(2022)07-0062-08 DOI:10.16526/j.cnki.11-4762/tp.2022.07.009 中图分类号:TP391 文献标识码:A

机场不正常事件实体检测与识别方法研究

侯启真，袁天一，王罗平
（中国民航大学 电子信息与自动化学院，天津 300300）

摘要：民航安全自愿报告系统收集的海量故障报告以非结构化文本形式存储，不便于相关人员针对大量不正常事件加以分析并采取控制措施；命名实体识别技术可以将海量非结构化文本中的关键要素进行检测和识别，抽取成类别分明的结构化信息，作为进一步分析不正常事件并加以控制的基础工作；将机场不正常事件报告作为研究对象，提出了一种基于神经网络的中文命名实体识别模型，对文本进行了结构化处理；针对随机选用的训练样本一些实体类别分布比较稀疏和人工标注费时费力的问题，提出了基于模型预测分数的样本选择策略，实现了预标注样本的高效筛选；经过实验验证，该模型与 BiLSTM _ CRF 模型、BiLSTM _ self-attention _ CRF 模型相比 F_1 值均提高了约 6 个百分点，该样本选择策略明显提高了人工标注效率，筛选出足够多的含有稀疏实体的样本。

关键词：命名实体识别；多尺度注意力；样本选择策略；双向长短时记忆网络；条件随机场

Research on Detection and Recognition Method of Airport Abnormal Event Entities

HOU Qizhen, YUAN Tianyi, WANG Luoping
(College of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China)

Abstract: The massive reports of fault events collected by the civil aviation safety voluntary reporting system are stored in the form of unstructured texts, which are not convenient for the relevant personnel to analyze and take the control measures for a large number of abnormal events. The technology of named entity recognition can detect and identify the key elements in the massive unstructured texts and extract them into the structured information with clear categories, which can be used as the foundation work for further analysis and control of abnormal events. As the reports of Airport abnormal events are taken as the research object, a neural network-based Chinese named entity recognition model based on Neural Network is proposed to structure the texts. For the problems of some entity categories sparse distribution of randomly selected training samples and time-consuming and laborious manual labeling, a sample selection strategy based on the model prediction scores is proposed to achieve the efficient screening of pre-labeled samples. After experimental validation, the model improves the F1 value by about 6 percentage points compared with the BiLSTM _ CRF model and the BiLSTM _ self-attention _ CRF model, and this sample selection strategy significantly improves the manual annotation efficiency, which screens out enough samples containing the sparse entities.

Keywords: named entity recognition; multi-scale self-attention mechanism; sample selection strategy; bi-directional long and short term memory network; conditional random field

0 引言

民航安全是民航业长久的主题[1]，在美国的航空安全自愿报告系统（ASRS，aviation safety reporting system）获得成功后，全世界众多国家纷纷开始建立适合自身实际的航空安全自愿报告系统，我国创建了中国民用航空安全自愿报告系统[2]。该系统所收集的报告中含有报告人所见所闻的民航安全隐患故障，需要总结归纳引发故障的原因和控制故障发生的措施来防止重大事故的发生，从而保障民航系统安全运行。随着时间积累，报告数量不断增长，每份报告的非结构文本所含要素信息得不到充分分析，传统的事件分析方法面对大量的文本很耗费人力也很依赖分析人员的专业能力。

为了充分利用这些事件报告，需要检测并提取出文本中的事件本质要素，这些要素存在于非结构化的文本中，且这些要素正是影响着民航运行安全的风险要素，主要是人、机、环境的一些状态信息。而命名实体识别正是能够做到检测和识别此类文本要素的关键技术，命名实体识别是一项序列标记任务，中文命名实体识别就是将每个文字或符号检测为其对应的实体类别。随着深度学习的兴起，循环神经网络（RNN，recurrent neural network）较适用于

收稿日期:2022-01-19； **修回日期**:2022-02-23。
基金项目:华东空管局科技项目(KJ2101)。
作者简介:侯启真(1966-),女,天津人,硕士,副教授,主要从事智慧机场、机场运行与保障技术等方向的研究。
通讯作者:袁天一(1995-),男,辽宁铁岭人,硕士研究生,主要从事自然语言处理、机场运行检测与控制技术方向的研究。
引用格式:侯启真,袁天一,王罗平.机场不正常事件实体检测与识别方法研究[J].计算机测量与控制,2022,30(7):62-69.

处理命名实体识别这样的序列标注任务[3]。但是面对长文本序列，RNN 的梯度消失与梯度爆炸的缺陷严重影响其序列标注效果。长短时记忆网（LSTM，long short-term memory）是一个特殊的循环神经网络，网络利用输入门、遗忘门和输出门来管理序列化数据[4-5]，在命名实体识别任务上取得了较为优异的效果。在此基础上有人提出双向长短时记忆网络（BiLSTM，Bi-directional long short-term memory）来提高模型效果，同时结合在命名实体识别任务上表现较好的机器学习模型——条件随机场（CRF，condition random fields），可以使得该任务在通用领域数据集上达到更好的识别效果。近几年也有人在此模型的基础上引入自注意力机制，在一定程度上提升了模型识别能力。

机场不正常事件是航空安全自愿报告中描述事件与机场相关的文本报告，经过人工筛选，并进行预处理得到命名实体识别模型需求的非结构化文本形式。机场不正常事件命名实体识别技术的任务是从非结构化的机场不正常事件文本中将该领域文本特定的不同类别实体检测识别出来，以达到对机场不正常事件关键要素提取和分类的目的，得到结构化文本作为开展机场不正常事件分析总结控制措施的基础工作。然而由于机场不正常事件文本在表述方式、事件状况、专业用语等文本特点上与通用领域不同，且通用领域主要以人名、地名、机构名等简单实体为命名实体识别目标，所以通用领域常用的命名实体识别模型在本领域很难达到较好效果。

因此，针对以上问题，提出了更适合于机场不正常事件文本数据的命名实体识别模型 BiLSTM _ MSA _ CRF（Bi-directional Long Short-Term Memory _ Multi-Scale Self-Attention _ Condition Random Fields）模型。此外，为降低人工标注成本，根据模型自身特点，设计了样本选择策略，在降低人工标注数据量的同时更高效地提高了模型泛化能力。

1　机场不正常事件报告的构造特征

机场不正常事件报告文本从整个文本角度，文本长度偏长，每份报告 300～700 字。上下文具有很强的相关性，长距离相关性将影响着命名实体识别效果。由于上下文的相关性也帮助丰富文本中关键要素的语义信息，使其明显区别于通用领域文本的结构，如“…27 号跑道发生跑道入侵事件，并未造成…”中“入侵”与“跑道”共同组合成一个词语“跑道入侵”有别于通用领域的常规用法，结合前文“27 号跑道”这一地点词可以确定此处词语语义。

从单个实体角度，文中含有一定量的专业性用语，中英文缩写及其中英文全称，以及中文、字母、数字多种字符串组合在文本中交替出现，这些字符串可能表达航路、航班、扇区等信息（例如 A326、SCS8997、ZSSSAR11），实体长度不等，实体间相互影响密切且交错。所需检测的实体种类也较多，多个实体种类之间比较相似，比如人的行为状态和其他生物的行为状态会有类似，需要结合语境进行区分。

2　数据标注规则

根据国际民航组织（ICAO）9859 号文件[6]，并结合机场不正常事件文本内容特点，充分考虑我国民航安全报告系统对故障防控的需求，设立了 14 个命名实体类别：时间、地点、方位、天气元素/能见度、航空器、航空器状态、航空器部件、航空器部件状态、设施、设施状态、人物类别、人类行为/状态、其他生物（不包括人类）、其他生物的状态。每个实体对应特定的编号，编号表如表 1 所示。

表 1　命名实体类别编号

实体类别	实体编号	实体类别	实体编号
时间	SJ	地点	DD
方位	FW	天气元素/能见度	TQ
航空器	HKQ	航空器状态	HKQZT
航空器部件	BJ	航空器部件状态	BJZT
设施	SS	设施状态	SSZT
人物类别	RL	人类行为/状态	RLZT
其他生物(不包括人类)	SW	其他生物的状态	SWZT

本文采用命名实体识别常用的 BIO 标注原则[7-8]对文本数据进行序列标注，即实体的开始标为 B，实体的非开头部分标为 I，非实体标为 O。由于每段文本较长，为方便人工标注，采用 {"*text*":"S","*label*":{e_1:[N_{e_1}],…,e_k:[N_{e_k}],…,$e_{|E|}$:[$N_{e_{|E|}}$]}} 标注方式，这种标注方式相对传统的 BIO 人工标注更简单便捷。其中，S 代表文本序列，$e_k \in E$ 是命名实体类别，N_{e_k} 代表在 S 这一文本序列中属于 e_k 这一实体类别的实体集合，人工标注完成的样本如图 1 所示。

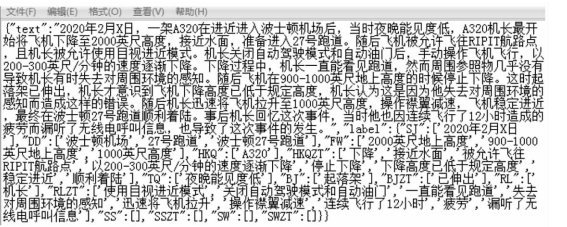

图 1　人工标注样本示例

数据处理程序中，将进行相应转换处理，程序经过如图 2 所示对人工标注数据进行相应处理，从而得到对应的 BIO 标注形式。

```
entity_bio_label = [0] * len(text)
for entity_type, entities in label.items():
    entity_type_code = args.entity_type[entity_type]
    for entity in entities:
        start_indexs = get_start_index(text, entity)
        for start_index in start_indexs:
            entity_bio_label[start_index] = entity_type_code * 2 - 1
            for i in range(start_index+1, start_index + len(entity)):
                entity_bio_label[i] = entity_type_code * 2
```

图 2　BIO 标注处理程序

3 命名实体识别方法和过程

依据各个领域现有命名实体识别模型[9-10]，并分析机场不正常事件报告的构造特征，提出的适用于检测机场不正常事件要素信息的命名实体识别任务，主要分为 4 个部分：文本向量化，双向长短时记忆网络和多尺度注意力机制（MSA，multi-scale self-attention）提取上下文特征信息以获取文本中每个字的实体类别预测分数，条件随机场将获取的最优预测序列解码输出最终识别结果，总体模型框架如图 3 所示。

3.1 字向量化

需要将输入的句子中每个字表示成字向量，字向量的表示方式主要分为两种：独热表示和稠密表示。由于独热表示无法表示字与字之间的相关关系，逐渐被新生的稠密表示方式取代，Word2vec[11]正是目前较经典的字向量稠密表示方法。Word2vec 可以表示字与字之间的相关关系，从而含有一定的语法和语义特征表示，进而从输入端提升命名实体识别模型的泛化能力。已知文本序列 $S=\{s_1,s_2,\cdots,s_m\}$ 有 m 个字，经过 Word2vec 处理后得到每个字 s_i 相对应的字向量表示形式 x_i，如式（1）所示：

$$x_i=\boldsymbol{W}^{w2v}v^i \tag{1}$$

其中：$\boldsymbol{W}^{w2v}\in R^{d_x\times|V|}$ 是由 Word2vec 训练得到的向量矩阵，d_x 是字向量的维度，$|V|$ 是输入字表的大小，v^i 是输入字 s_i 的词袋表示（独热形式）。由此得到一个向量序列 $x=\{x_1,x_2,\cdots,x_m\}$，作为命名实体识别网络的字向量输入。

3.2 提取上下文信息

单向 LSTM 可随着序列信息的提取保留前文“值得记忆”的特征信息，而模型最后检测出的序列标签是结合前文的信息预测得出的，也就做到了结合上文的语境信息来做命名实体识别任务。为解决 RNN 在长文本序列标注任务上的缺陷，每个 LSTM 均包含着输入门、遗忘门和输出门这 3 个“门”单元结构，以降低梯度消失等问题的出现率。LSTM 单元结构如图 4 所示。

式（2）描述了 LSTM 具体计算过程。

$$\begin{aligned}
i_t&=\sigma[\boldsymbol{W}_i\cdot(h_{t-1},x_t)+b_i]\\
f_t&=\sigma[\boldsymbol{W}_f\cdot(h_{t-1},x_t)+b_f]\\
o_t&=\sigma[\boldsymbol{W}_o\cdot(h_{t-1},x_t)+b_o]\\
\tilde{c}_t&=\tanh[\boldsymbol{W}_c\cdot(h_{t-1},x_t)+b_c]\\
c_t&=f_t\odot c_{t-1}+i_t\odot\tilde{c}_t\\
h_t&=o_t\odot\tanh[c_t]
\end{aligned}\tag{2}$$

其中：σ 是 sigmoid 函数，tanh 为双曲正切函数；$\boldsymbol{W}_i$，$\boldsymbol{W}_f$，$\boldsymbol{W}_o$，$\boldsymbol{W}_c$ 分别是输入门、遗忘门、输出门和更新细胞状态的权重矩阵，b_i，b_f，b_o，b_c 为对应的偏置。首先，以 $t-1$ 时刻的隐藏层状态 h_{t-1} 和当前 t 时刻的字向量 x_t 为输入，分别计算出输入门值 i_t，遗忘门值 f_t，临时细胞状态 $\tilde{c}_t$。以此为基础，结合 $t-1$ 时刻的细胞状态 c_{t-1} 计算出当前的细胞状态值 c_t。然后，在计算输出门值 o_t 的基础上，得到当前时刻的隐藏层输出 h_t。

为获得上下文语境的信息，双向长短时记忆网络（BiLSTM）被 Alex Graves[12]提出，从而解决了 LSTM 没有充分利用下文信息的缺点，也刚好符合人类分析文章信息的方法，即结合上下文分析字、词以及词组所表达的语义倾向，BiLSTM 也是当前最常用的命名实体识别模型的一部分，取代了 RNN 和 LSTM。前后两个方向的 LSTM 组成了 BiLSTM 网络，如图 3 所示，网络结构包括两个子网络，包括正向和反向的上下文信息[13]，因此对于第 i 个字向量 x_i 经过 BiLSTM 模块后得到的隐藏层向量表示 h_i 由正向网络和反向网络输出的向量 $\overrightarrow{h_i}$ 和 $\overleftarrow{h_i}$ 得到：

$$h_i=\overrightarrow{h_i}\oplus\overleftarrow{h_i} \tag{3}$$

字向量经过 BiLSTM 提取一定的上下文特征，但并不足以准确检测每个字的对应标签。

3.3 结合层次结构的自注意力机制

尽管双向长短时记忆网络在一定程度上已经保留了上下文“重要”信息，已经可以做到较全面的处理，但是其依然没有对这些“重要”信息分清主次，即从 BiLSTM 中得到每个字向量对应的上下文特征向量，但并没有考虑到不同词语间的不同程度关系，也没有充分考虑到不同的词语对模型识别结果会产生不同程度的影响，所以识别效果更待提升，需要使用自注意力机制来帮助分配权重以解决此问题。所以结合了自注意力机制的命名实体识别模型更能够提取更加主要且与现有输出关联度更高的特征信息，避免过多提取次要关联信息而造成语义偏差，在对输入向量施加合适的权重系数后，模型识别结果会得到有效

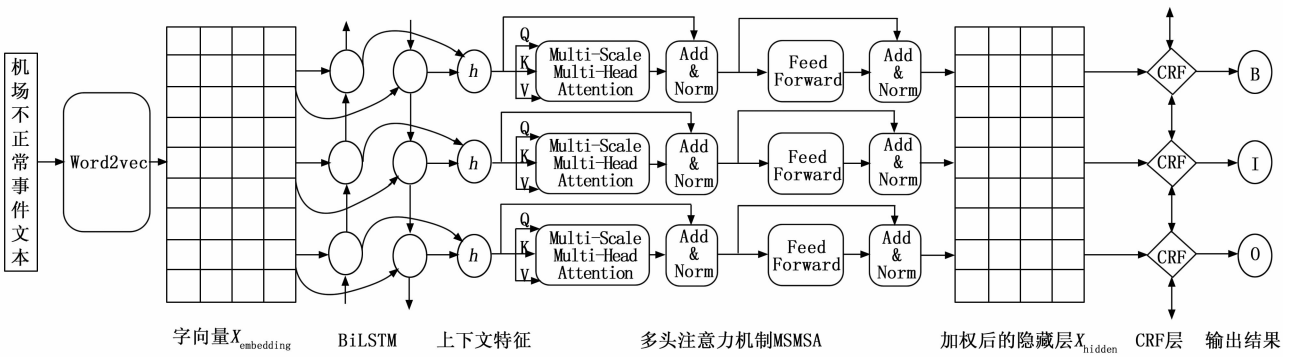

图 3 机场不正常事件命名实体识别的 BiLSTM-MSA-CRF 模型构架

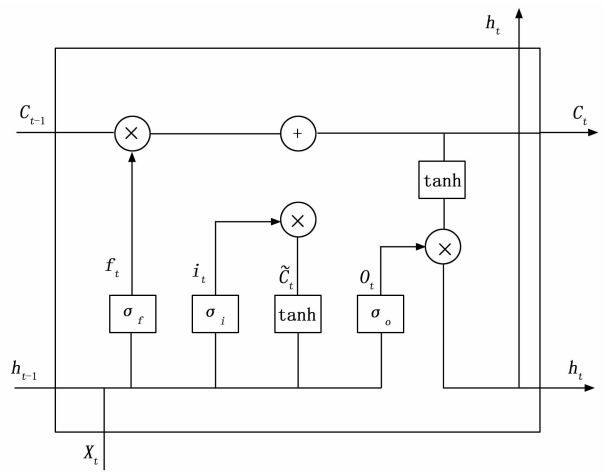

图 4　LSTM 单元结构

提升。

近年来，诸多领域为解决命名实体识别的这一问题引入了自注意力机制[14-16]，尽管自注意力可以建模非常长的依赖关系，但深层的注意力往往过度集中在单个字上，且权重过于分散，并不能构成词语间的依赖关系，导致对局部信息的使用不足，对短序列自注意力相对有效，但其难以表示长序列，随着句子的长度增加自注意力的性能逐渐下降，从而导致信息表达不足，给模型完整地理解数据信息带来困难，在语境中应更主要以词与词之间的影响来作为特征，这样才更能提高模型识别效率。且基于自注意力机制的方法缺乏先验假设，需要很大的样本数据集才能训练出一个泛化能力较好的模型。本研究数据量有限，无法满足大样本数据集的要求。多尺度结构可以帮助模型捕捉不同尺度的特征，实现多尺度的常用方法是采用层次结构，通过层次结构，模型可以捕获较低层次的局部特征和较高层次的全局特征。多尺度多头注意力[17]的各个头具有可变尺度，头部的大小限制了自注意力的工作范围：大尺度包含更多上下文信息，小尺度更关注局部信息。

BiLSTM 输出向量为 h_i，对应的序列矩阵为 $\boldsymbol{H}=\{h_1, h_2, \cdots, h_n\}$，其中 $\boldsymbol{H}\in R^{n\times D}$，$n$ 为句子长度，D 是 h_i 的向量维度，式（4）描述了多尺度注意力的计算过程。

$$head_j(\boldsymbol{H},\omega_j)_i = softmax\left[\frac{Q_{ij}C_{ij}(K,\omega_j)^T}{\sqrt{D}}\right]C_{ij}(V,\omega_j)$$

$$C_{ij}(A,\omega_j)=\{A_{i-\omega_j-1/2,j},\cdots,A_{i+\omega_j-1/2,j}\}$$

$$\boldsymbol{Q}=\boldsymbol{H}\cdot\boldsymbol{W}^Q,\boldsymbol{K}=\boldsymbol{H}\cdot\boldsymbol{W}^K,\boldsymbol{V}=\boldsymbol{H}\cdot\boldsymbol{W}^V$$

$$head_j(\boldsymbol{H},\omega_j)=Concat[head_j(\boldsymbol{H},\omega_j)_1,\cdots,head_j(\boldsymbol{H},\omega_j)_n]$$

$$MSMSA(\boldsymbol{H},\Omega)=Concat[head_1(\boldsymbol{H},\omega_1),\cdots, head_j(\boldsymbol{H},\omega_j),\cdots,head_{N'}(\boldsymbol{H},\omega_{N'})]\boldsymbol{W}^O \quad (4)$$

其中：$\boldsymbol{W}^Q$，$\boldsymbol{W}^K$，$\boldsymbol{W}^V$，$\boldsymbol{W}^O$ 是可学习的参数矩阵，ω 是每个头的尺度大小，ω_j 即为第 j 个头的尺度，共有 N' 个头，多尺度多头自注意力的所有头的尺度集合为 $\Omega=[\omega_1,\cdots,\omega_j,\cdots,\omega_{N'}]$，$C$ 为给定位置提取上下文特征的函数。

多头多尺度注意力机制，在不同层分配了不同尺度的“头”，不同层中对应的尺度分配遵循式（5）：

$$z_k^l=\begin{cases}0 & l=L \quad or \quad k=|\Omega| \\ z_{k+1}^l+\dfrac{\alpha}{l} & k\in\{0,\cdots,|\Omega|-1\}\end{cases}$$

$$n_k^l = softmax(z_k^l)\cdot N' \quad (5)$$

其中：L 表示多尺度多头注意力机制的层数，$|\Omega|$ 表示候选尺度大小的个数，z_k^l 是中间变量，n_k^l 表示第 l 层、第 k 个尺度大小的头部个数，N' 是该层总头数，α 表示控制每一层尺度分配的超参数。

以上公式计算过程对于单个向量 h_i 可以归结为式（6）：

$$h'_i=\sum_{j=1}^{n}\alpha_{ij}\cdot h_j \quad (6)$$

对应注意力计算结构图如图 5 所示。

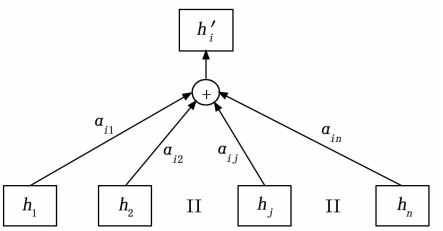

图 5　注意力加权计算过程

其中：h'_i 表示利用多尺度自注意力机制输出新的字特征向量，它是由 BiLSTM 模型输出的各特征向量 h_j 与对应的权重 α_{ij} 的乘积求和计算得到。采用多尺度自注意力权重分配方法来改变双向 LSTM 输出的概率矩阵，可以兼顾更多局部特征，也就能改善 CRF 层的序列标注结果。

3.4　解码输出检测结果

CRF[18]解码过程中，将重新分配权重后的双向 LSTM 概率矩阵输出结果作为输入，获得预测序列标签。CRF 模型关注输入序列各个相邻字的前后依赖关系，进而计算最优预测标签序列。借鉴王栋[19]等人使用 CRF 模型的思路，相关公式计算过程如下：

记句子序列为 $S=\{s_1,s_2,\cdots,s_m\}$，其预测的标签序列为 $Y=\{y_1,y_2,\cdots,y_m\}$，则序列预测得分矩阵计算如式（7）：

$$Score(S,Y)=\sum_{i=1}^{n}(T_{y_{i-1},y_i}+P_{i,y_i}) \quad (7)$$

其中：$\boldsymbol{T}$ 代表状态转移矩阵，T_{y_{i-1},y_i} 为 y_{i-1} 标签转移到 y_i 标签的概率得分，P_{i,y_i} 是第 i 个字符被标记为标签 y_i 的概率得分。文本序列 S 计算产生标记序列 Y 的概率如式（8）所示：

$$p(Y\mid S)=\frac{e^{Score(S,Y)}}{\sum_{\tilde{Y}\in Y^s}[e^{Score(S,\tilde{Y})}]} \quad (8)$$

在训练过程的标记序列的似然函数如式（9）所示，通过极大似然估计的方法估计条件随机场的模型参数。

$$\log[p(Y\mid S)]=Score(S,Y)-\log\{\sum_{\tilde{Y}\in Y^s}[e^{Score(S,\tilde{Y})}]\} \quad (9)$$

使用 CRF 对序列进行预测时利用维特比（Viterbi）算法求解最可能的序列标签，最终输出如式（10）所示的最

优序列 Y^*。

$$Y^* = \underset{\tilde{Y}\in Y^S}{argmax}[Score(S,\tilde{Y})] \tag{10}$$

其中：$\tilde{Y}$ 是真实序列，Y^S 是所有标记集合。

3.5 样本选择策略

由于模型所需标注训练样本数量较大，人工标注成本较高，且已有训练数据中各个类别的实体数量不均衡，以至出现比较稀疏的实体类别，从而导致模型对这些稀疏实体识别不准确，为检测出含有此类实体的高质量训练样本和提高人工标注效率，本文根据数据和模型本身特点，设计了基于不确定性的样本选择策略。该方法既能减低人工标注成本又能更高效地提高模型的泛化能力，基于不确定性的样本选择策略的核心思想是模型无法进行有效判断的样本[20-22]。结合现有命名实体识别模型，本文使用最优预测序列概率 $p(Y^*|S)$ 作为模型对未标注样本的不确定性评判依据，最优预测序列概率 $p(Y^*|S)$ 越低，模型对样本序列的标注越不确定，这类样本与已有训练数据相比含有稀疏实体较多，这类样本越值得加入训练集。基于不确定性的样本选择策略如式（11）。

$$D(Y^*) = \{Y^* \mid p(Y^*|S) \leqslant P_D\} \tag{11}$$

其中：$D(Y^*)$ 是通过选择后得到的需人工标注的样本集，P_D 为模型最优预测序列概率阈值，当样本 S 对应的最优预测序列 Y^* 的概率未达到阈值时，则将该样本加入需人工标注样本集，等待人工进行标注。使用该样本选择策略后，构成了与模型训练模块构成了闭环主动学习框架，如图 6 所示。

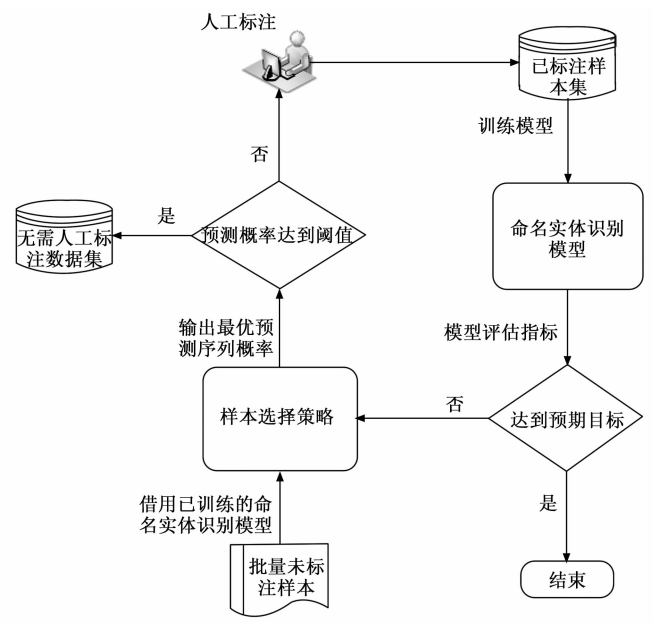

图 6 融合样本选择策略的命名实体识别框架

4 实验结果与分析

4.1 实验数据准备

使用的数据来自于 ASRS 和中国民用航空安全自愿报告系统中与机场相关的航空安全自愿报告，选取的报告包含了 2010～2021 年间机场航空安全自愿报告 10 536 条，所有文本去除无效字符并整理格式后组成本实验机场不正常事件样本数据，数据以中文形式呈现，每篇报告 500 字左右。随机选取了 7 000 条样本进行人工标注，标注形式如图 1 所示，并随机将其分为 5 000 条文本的训练集和 2 000 条文本的测试集。剩余的未标注样本作为样本选择策略的实验数据。

4.2 实验环境、参数设置和评价指标

实验在 Windows10（64 位）系统中使用 Python3.6 作为编程语言，基于 Pytorch 框架对本文方法和对比实验方法进行程序实现。所有实验是在 Intel Core i7-8700 处理器、16 G 内存、NVIDIA Quadro P2000 GPU 硬件设备条件下进行的。表 2 是实验中模型参数设置情况。

表 2 模型参数设置

参数	数值
Embedding_size	300
Hidden_size	150
Dropout	0.5
Learning_rate	0.000 1
Epoch	50
Early_stop	10
Embedding_size	300
Hidden_size	150

实验采用精确率 P、召回率 R 和 F_1 值对命名实体识别结果进行评价。3 个评价指标的计算如下：

$$P = \frac{\text{识别正确的实体数目}}{\text{识别出实体的总数目}}$$

$$R = \frac{\text{识别正确的实体数目}}{\text{样本实际实体实体的总数目}}$$

$$F_1 = \frac{2\times P\times R}{P+R} \tag{12}$$

以下实验均通过计算不同模型在相同数据上的精确率 P、召回率 R 和 F_1 值进行对比。

4.3 实验结果对比与分析

实验一：加入多尺度注意力机制的命名实体识别模型在机场不正常事件文本数据上的识别效果需要对比通用领域的常用方法来验证，以证明多尺度注意力机制能够改善机场不正常事件文本命名实体识别效果。实验使用 3.1 节所提及的训练集和样本集分别训练 BiLSTM_CRF 模型、BiLSTM_self-attention_CRF 模型以及本文提出的 BiLSTM_MSA_CRF 模型，为降低选取数据的偶然性，经过 5 次随机分配得到的训练数据和测试数据来分别训练 3 个模型，最终将模型得出的评价指标取平均值，并填写入表 3 中。

从表 3 可以看出，加入自注意力机制后，模型识别能力确有提升，但并不明显，这正是因为自注意力机制对于文段较长且识别结果很依赖上下文语境的文本并没有很好地发挥其捕捉上下文重要信息的作用，注意力过于分散在单个字

上，没有充分利用词语级别的局部信息。而加入多尺度注意力机制后，识别效果有了明显提升，说明多尺度注意力能够改善自注意力的缺点，更适合机场不正常事件这种长文段的命名实体识别。

表 3　固定样本集条件下不同模型的对比实验结果

模型	P/%	R/%	F1/%
BiLSTM_CRF	83.26	84.83	84.04
BiLSTM_self-attention_CRF	84.53	85.33	84.93
BiLSTM_MSA_CRF	88.68	92.63	90.61

为了降低人工标注成本且更高效地提升模型泛化能力，使用 2.6 节提出的样本选择策略进行对比实验。如图 7 所示，是阈值设为 0.9 时，提示需要标注的样本示例。

```
预测实体为: [{'entity_text': '中心管制员', 'start_pos': 0, 'end_pos'
0.8678243892533438
建议标注此样本.....................................
```

图 7　样本选择策略下程序提示需要标注的样本示例

实验二：分别对比不同概率阈值 P_D 对 3 种命名实体识别模型的影响，以寻找一个更合适的阈值。

实验步骤为：将前一次实验训练后的 3 种模型保存分别命名为 BiLSTM_CRF、BiLSTM_self-attention_CRF 和 BiLSTM_MSA_CRF，选取 4 种不同最优预测序列概率阈值 P_D（分别为 0.8、0.85、0.9、0.95），并分为 4 个批次逐渐增加选取样本，每个批次随机选取 500 条未标注样本，3 种模型经样本选择策略后，挑选未达阈值的样本进行人工标注，加入训练集进行模型再训练，将不同阈值各个批次训练完成的模型区分开命名 BiLSTM_CRFm(n)，其中 $m=[0.8,0.85,0.9,0.95]$ 为阈值，$n=[1,2,3,4]$ 为挑选样本批次，如 BiLSTM_CRF0.8(1)代表设定阈值为 0.8 随机选取 500 条未标注样本，筛选出需要人工标注的样本加入训练集中对 BiLSTM_CRF 再训练而得到的模型；BiLSTM_MSA_CRF0.9(3)代表设定阈值为 0.9 在未标注样本集里随机选取 500 条未标注样本筛选出需要人工标注的样本累加到训练集中对 BiLSTM_MSA_CRF 再训练而得到的模型。对不同的样本需要标注的内容有一定的差异性，为防止因这种“参差不齐”的现象而引起的偏差，在所有未标注样本中随机进行 5 次抽选 2 000 条文本，各进行 5 次实验取平均值作为最终实验结果。实验结果如图 8～13 所示，每组的横坐标为使用样本选择策略选取备选样本的批次，即 500(2)意为第二批次随机拿出 500 条未标注样本进行样本选择，图 8，10，12 纵坐标为每个批次筛选出需要人工标注的样本数量，图 9，11，13 纵坐标为对应批次训练后模型的 F_1 值。

由图 8～13 以及表 4 可以看出，随着样本选择策略的使用，3 个模型的精确率 P、召回率 R 和 F_1 值均有提升，并且随着阈值的提升（由 0.8 到 0.85 再到 0.9）模型的精确率 P、召回率 R 和 F_1 值在以更高的增长速率提升，并且增长趋势有

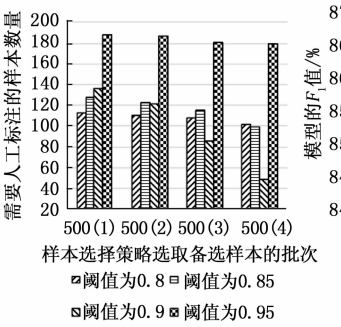

图 8　不同阈值条件下 BiLSTM_CRF 模型需人工标注样本量对比

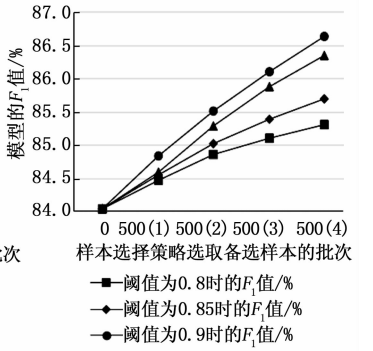

图 9　不同阈值条件下 BiLSTM_CRF 模型随人工标记轮次 F_1 值变化情况

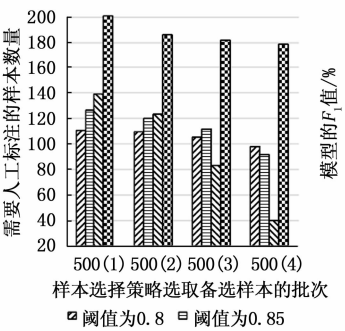

图 10　不同阈值条件下 BiLSTM_self-attention_CRF 模型需人工标注样本量对比

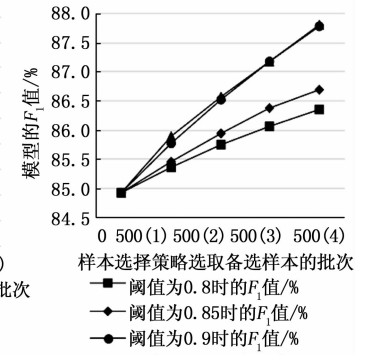

图 11　不同阈值条件下 BiLSTM_self-attention_CRF 模型随人工标记轮次 F_1 值变化情况

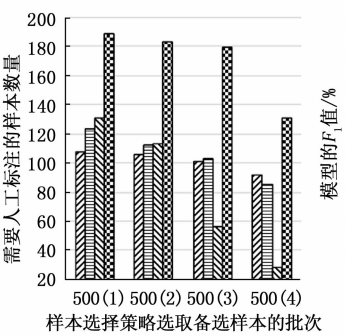

图 12　不同阈值条件下 BiLSTM_MSA_CRF 模型需人工标注样本量对比

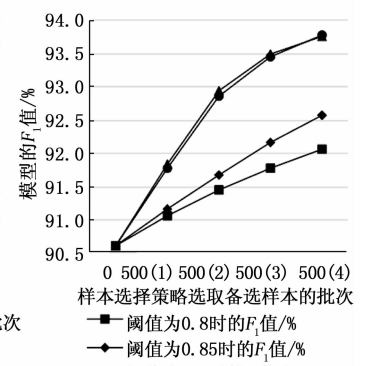

图 13　不同阈值条件下 BiLSTM_MSA_CRF 模型随人工标记轮次 F_1 值变化情况

提前趋于平稳的趋势，这是因为各个模型对大部分能够准确识别的未标注样本的预测分数主要集中在 0.9 以上，预测分数在 0.9 之下的样本正是模型不确定性较高的样本，需要加入训练集来提升模型的泛化能力。此外，随着阈值的提升（由 0.8 到 0.85 再到 0.9）模型所需标注样本量也跟随着筛

表 4 不同阈值条件下 3 种模型评价指标变化对比

模型	未标注样本筛选批次	阈值为 0.8			阈值为 0.85			阈值为 0.9			阈值为 0.95		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
BiLSTM_CRF	500(1)	83.77	85.20	84.48	83.88	85.24	84.55	84.19	85.48	84.83	84.23	84.96	84.59
	500(2)	84.16	85.55	84.85	84.41	85.63	85.02	85.01	86.01	85.51	84.93	85.66	85.29
	500(3)	84.37	85.86	85.11	84.80	85.99	85.39	85.74	86.48	86.11	85.66	86.13	85.89
	500(4)	84.47	86.15	85.30	85.13	86.27	85.70	86.40	86.89	86.64	86.21	86.52	86.36
BiLSTM_self-attention_CRF	500(1)	85.05	85.69	85.37	85.17	85.76	85.46	85.50	86.04	85.77	85.66	86.12	85.89
	500(2)	85.47	86.02	85.74	85.73	86.17	85.95	86.38	86.66	86.52	86.49	86.67	86.58
	500(3)	85.80	86.34	86.07	86.20	86.56	86.38	87.17	87.21	87.19	87.15	87.22	87.18
	500(4)	86.06	86.64	86.35	86.51	86.87	86.69	87.87	87.71	87.79	87.93	87.68	87.80
BiLSTM_MSA_CRF	500(1)	83.77	85.20	84.48	83.88	85.24	84.55	84.19	85.48	84.83	84.23	84.96	84.59
	500(2)	84.16	85.55	84.85	84.41	85.63	85.02	85.01	86.01	85.51	84.93	85.66	85.29
	500(3)	84.37	85.86	85.11	84.80	85.99	85.39	85.74	86.48	86.11	85.66	86.13	85.89
	500(4)	84.47	86.15	85.30	85.13	86.27	85.70	86.40	86.89	86.64	86.21	86.52	86.36

选批次逐渐减少，阈值为 0.9 时现象尤为明显。阈值为 0.9 和 0.95 时 3 个模型的评价指标上升趋势均几乎重合，模型所需标注样本量却有明显差异，模型预测分数能达到 0.95 的样本近乎少数，所以阈值设为 0.95 时 3 个模型需人工标记的样本量明显多于阈值设为 0.9 的情况，不过在阈值为 0.95 时，BiLSTM _ MSA _ CRF 模型随着样本选择策略批次所需人工标记的样本量下降速度更明显些，也从一定程度上说明该模型预测分数高于 0.95 的样本数量要比另两种模型多。所以 4 个阈值相比，阈值 0.9 更适合作为本文的文本数据和本文所使用的命名实体识别模型。

实验三：为了更加凸显样本选择策略的作用，在不使用样本选择策略的情况下，在上一个实验的同一批 500 条样本中随机选出与该批样本选择策略选出的样本数目相同的未标注样本加入训练集训练相应模型，实验结果如图 14 所示。

从图 14 可以看出，在未使用样本选择策略的情况下，人工标注与阈值为 0.9 的样本选择策略相同数量的样本，模型识别能力的提升效果很不明显，与使用了样本选择策略差异很大，所以样本选择策略明显帮助我们在一批样本中检测出更能提升模型泛化能力的“有用”样本，人工标注后加入训练集，帮助模型“查漏补缺”。

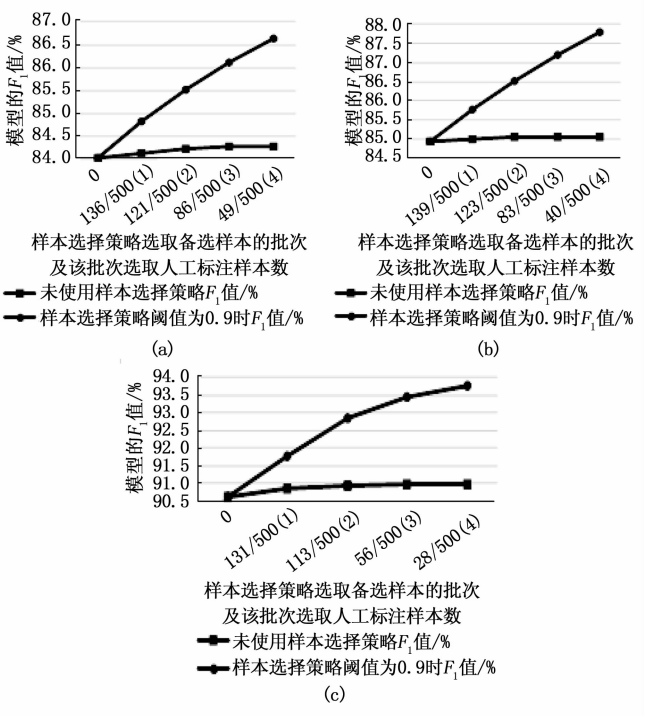

图 14 BiLSTM _ CRF、BiLSTM _ self-attention _ CRF、BiLSTM _ MSA _ CRF 使用和未使用样本选择策略实验结果对比

5 结束语

经过上述 3 个实验的对比，在机场不正常事件数据上，本文提出的 BiLSTM _ MSA _ CRF 模型达到更好的识别效果，明显比 BiLSTM _ CRF、BiLSTM _ self-attention _ CRF 提升了 6 个百分点的 F_1 值。样本选择策略降低了人工标注成本，且帮助模型挑选了含有稀疏实体的样本来供给人工标注后加入训练数据，实验得出的 F_1 表明该方法明显提升了模型识别效果。实验证明本文提出的方法是解决海量机场不正常事件的关键要素检测和识别的有效方法，可作为进一步分析大量机场不正常事件文本的基础工作，协助民航相关人员及时总结事故规律和关系、制定控制事故的措施。

参考文献：

[1] ICAO, Annex 19-safety management [Z]. 2016.

[2] 刘俊杰，杜尹岚，闫慧娟. Python 环境下的航空安全报告信息分析方法 [J]. 科学技术与工程，2021，21 (10)：4278 - 4283.

[3] 邓博研. 面向工业领域知识图谱构建的信息抽取方法研究[D]. 广州：广东工业大学，2020.
[4] 姜同强，王岚熙. 基于双向编码器表示模型和注意力机制的食品安全命名实体识别[J]. 科学技术与工程，2021，21 (3)：1103－1108.
[5] 高　凡，李　樊，张　铭，等. 基于文本挖掘的高速铁路动车组故障多级分类研究[J]. 计算机测量与控制，2020，28 (7)：59－63.
[6] Doc 9859，International Civil Aviation Organization (ICAO). Safety Management Manual [Z]. 2018.
[7] ERIK F. Tjong Kim Sang，Fien De Meulder. Introduction to the CoNLL-2003 Shared Task：Language-Independent Named Entity Recognition [J]. CoRR，2003，cs. CL/0306050.
[8] 郑　达. 稳健对话系统关键技术研究[D]. 上海：上海交通大学，2017，1－6.
[9] 孙　弋，梁兵涛. 基于BERT和多头注意力的中文命名实体识别方法[J/OL]. 重庆邮电大学学报（自然科学版)：1－10. [2022－01－19]. http://kns. cnki. net.
[10] 李鸿飞，刘盼雨，魏　勇. 基于自注意力和Lattice-LSTM的军事命名实体识别[J]. 计算机工程与科学，2021，43 (10)：1848－1855.
[11] YOAV GOLDBERG，OMER LEVY. word2vec Explained：deriving Mikolov et al. 's negative-sampling word-embedding method [J]. CoRR，2014，3722：1－5.
[12] ALEX GRAVES，JüRGEN SCHMIDHUBER. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks，2005，18 (5)：602－610.
[13] 李　丹，贾桂敏，程方圆，等. 陆空通话复诵语义自动化校验BiLSTM模型[J]. 信号处理，2019，35 (1)：57－64.
[14] 罗　熹，夏先运，安　莹，等. 结合多头自注意力机制与BiLSTM-CRF的中文临床实体识别[J]. 湖南大学学报（自然科学版)，2021，48 (4)：45－55.
[15] SHUFENG HE，DIANQI SUN，ZHAO WANG. Named entity recognition for Chinese marine text with knowledge-based self-attention [J]. Multimedia Tools and Applications，2021 (prepublish)：1－15.
[16] WAN Q，LIU J，WEI LN，et al. A self-attention based neural architecture for Chinese medical named entity recognition [J]. Math Biosci Eng，2020，17 (4)：3498－3511.
[17] QIPENG GUO，XIPENG QIU，PENGFEI LIU，et al. Multi-Scale Self-Attention for Text Classification [J]. Proceedings of the AAAI Conference on Artificial Intelligence，2020，34 (5)：7847－7854.
[18] XUEZHE MA，EDUARD H. Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF [J]. CoRR，2016，abs/1603.01354：1064－1074.
[19] 王　栋，李业刚，张　晓，等. 基于准循环神经网络的中文命名实体识别[J]. 计算机工程与设计，2020，41 (7)：2038－2043.
[20] 刘　畅. 基于主动学习和半监督机制的偏标记问题研究[D]. 保定：河北大学，2021.
[21] 张宏涛. 面向生物文本的实体关系自动抽取问题研究[D]. 北京：清华大学，2012.
[22] 王莉莉，付忠良，陶　攀，等. 基于主动学习不平衡多分类AdaBoost算法的心脏病分类[J]. 计算机应用，2017，37 (7)：1994－1998.

(上接第 61 页)

[11] 陈诗乐，王　笑，周昌军. 基于GA-Transformer模型的多因子股票预测[J]. 广州大学学报（自然科学版)，2021，20 (1)：44－55.
[12] 李　文，邓　升，段　妍，等. 时间序列预测与深度学习：文献综述与应用实例[J]. 计算机应用与软件，2020，37 (10)：64－70.
[13] 刘　甲，孙德山. 基于注意力机制和LSTM网络的股价预测[J]. Advances in Applied Mathematics，2021 (10)：4379.
[14] 王　彤. 基于Z-Stack协议栈的ZigBee网络组网研究与实现[D]. 保定：河北大学，2012.
[15] 高　翔，邓永莉，吕愿愿，等. 基于Z-Stack协议栈的ZigBee网络节能算法的研究[J]. 传感技术学报，2014，27 (11)：1534－1538.
[16] LI J，HU Y. Design of ZigBee network based on CC2530 [J] [J]. Electronic design engineering，2011，19 (16)：108－111.
[17] 李俊斌，胡永忠. 基于CC2530的ZigBee通信网络的应用设计[J]. 电子设计工程，2011，19 (16)：108－111.
[18] 景　强. 基于CC2538无线传感器网络节点设计研究[D]. 太原：中北大学，2018.
[19] 曾宝国. Z-STACK协议栈应用开发分析[J]. 物联网技术，2011，1 (3)：71－73.
[20] 李　军，黄　岚，王忠义. 基于Z-Stack协议栈的WSN能量管理策略[J]. 计算机工程，2011，37 (7)：121－124.
[21] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [C] // Advances in neural information processing systems，2017，30.
[22] CHOLAKOV R，KOLEV T. Transformers predicting the future. Applying attention in next-frame and time series forecasting [Z]. arXiv preprint arXiv：2108.08224，2021.
[23] TANG B，MATTESON D. Probabilistic Transformer For Time Series Analysis [J]. Advances in Neural Information Processing Systems，2021，34.
[24] ZHANG T，SONG S，LI S，et al. Research on gas concentration prediction models based on LSTM multidimensional time series [J]. Energies，2019，12 (1)：161.
[25] TAKASE S，OKAZAKI N. Positional encoding to control output sequence length [Z]. arXiv preprint arXiv：1904.07418，2019.
[26] 王　鑫，吴　际，刘　超，等. 基于LSTM循环神经网络的故障时间序列预测[J]. 北京航空航天大学学报，2018，44 (4)：772－784.
[27] CARON M，MÜLLER O. Hardening Soft Information：A Transformer-Based Approach to Forecasting Stock Return Volatility [C] //2020 IEEE International Conference on Big Data (Big Data). IEEE，2020：4383－4391.