郑佳明, 沈颖, 刘晓强, 等. 基于机器阅读理解的行车故障诊断知识抽取[J]. 智能计算机与应用,2024,14(9):56-62. DOI: 10.20169/j.issn.2095-2163.240908

基于机器阅读理解的行车故障诊断知识抽取

郑佳明[1,2], 沈 颖[2], 刘晓强[1], 涂文奇[1], 李柏岩[1]
(1 东华大学 计算机科学与技术学院, 上海 201620; 2 上海市计算机软件评测重点实验室, 上海 201112)

摘 要: 行车故障调查单是对行车故障诊断过程的文本记录,基于这些历史记录构建知识图谱可以更好地支持行车故障诊断智能化。由于该语料具有实体嵌套、实体跨度大、关系重叠等特点,传统的命名实体识别和关系抽取模型难以对其进行有效的知识抽取。针对语料中存在的实体嵌套和长实体识别问题,本文提出了一种融合强化学习的机器阅读理解模型,以问答形式进行实体识别,以指针网络进行解码;对于语料中存在的关系重叠问题,将关系抽取分为先识别主体再识别客体的两阶段,将不同实体对的关系抽取进行隔离。实验结果表明,基于机器阅读理解的方法在行车故障诊断领域的知识抽取上具有较好的性能,可以有效支持领域知识图谱构建。

关键词: 行车故障诊断; 知识图谱; 知识抽取; 机器阅读理解; 指针网络
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 2095-2163(2024)09-0056-07

Knowledge extraction of crane fault diagnosis based on machine reading comprehension

ZHENG Jiaming[1,2], SHEN Ying[2], LIU Xiaoqiang[1], TU Wenqi[1], LI Baiyan[1]
(1 School of Computer Science and Technology, Donghua University, Shanghai 201620, China;
2 Shanghai Key Laboratory of Computer Software Testing & Evaluation, Shanghai 201112, China)

Abstract: The crane fault investigation form is a text record of the crane fault diagnosis process. Constructing a knowledge graph based on these historical records can better support intelligent crane fault diagnosis. However, due to the characteristics of nested entities, large entity spans, and overlapping relations in this corpus, traditional named entity recognition and relationship extraction models are unable to perform effective knowledge extraction. To address the problems of nested entities and long entity recognition, this paper proposed a machine reading comprehension model fused with reinforcement learning. The model performed entity recognition in a question-answer format and decoded the output using a pointer network. For the problems of overlapping relations, relationship extraction was divided into two stages: first recognizing the subject and then recognizing the object, to isolate the relationship extraction of different entity pairs. Experiments show that the machine reading comprehension method has good performance in knowledge extraction for crane fault diagnosis, and can effectively support the construction of domain knowledge graphs.

Key words: crane fault diagnosis; knowledge graph; knowledge extraction; machine reading comprehension; pointer network

0 引 言

行车即起重机,是现代工业领域经常用到的起重设备,包含多部件的复杂自动化机械设备,长时间高强度的工作会导致故障发生,而目前故障诊断主要依靠人工经验,运维效率低下[1]。虽然历史故障调查单记录了每次故障的处理经过、故障原因、解决措施等信息,但这些信息往往以半结构化或非结构化的形式存储,后期检索和利用困难。因此,对故障调查单信息进行有效抽取,构建行车故障诊断知识图谱,可以充分利用历史数据以提升故障诊断效率[2]。

作者简介: 郑佳明(1998-),男,硕士研究生,主要研究方向:自然语言处理,知识图谱; 沈 颖(1974-),女,硕士,高级工程师,主要研究方向:软件质量,智能检测;涂文奇(1997-),男,硕士研究生,主要研究方向:自然语言处理,NL2SQL; 李柏岩(1968-),男,博士,副教授,硕士生导师,主要研究方向:机器学习,信号处理。
通讯作者: 刘晓强(1968-),女,博士,教授,硕士生导师,主要研究方向:智能信息处理,知识管理。Email:liuxq@dhu.edu.cn

收稿日期: 2023-05-16

知识抽取是构建知识图谱的关键步骤，可以分为两个子任务：命名实体识别和关系抽取。命名实体识别是自然语言处理领域中的基础任务，其目的是从文本中识别出特定类别的实体，根据实体特点又可以分为嵌套实体和扁平实体[3]。目前主流的扁平实体识别方法是将实体识别作为序列标注任务，使用 BIO（Begin－Intermediate－Other）或 BIOES（Begin－Intermediate－Other－ End－ Single）等方法进行序列标注，常用的模型一般是 BERT（Bidirectional Encoder Representations from Transformers）－BiLSTM（Bi－directional Long Short－Term Memory）－CRF（Conditional Random Fields）[4]。嵌套实体由于单个字可能存在多个标签，因此传统序列标注模型难以对其进行识别[5]。近年来出现了针对嵌套实体识别的研究，如 Ju 等[6]堆叠多层网络模型，根据嵌套深度动态进行实体识别，初步对嵌套实体的识别进行了尝试；Shibuya 等[7]通过次优序列学习，由外向内迭代识别嵌套实体，在嵌套实体的识别上取得了良好的效果。关系抽取作为自然语言处理领域的级联任务，旨在识别文本中实体间的关系，主要有流水线方法和联合抽取方法[8]。流水线方法是指先从文本中识别实体对，再对实体对进行关系分类[9]；联合抽取方法是指通过参数共享、联合解码等方式同时进行实体关系抽取[10]，如 Zheng 等[11]提出一种标注策略，标注实体的同时标注关系，通过联合解码进行实体关系联合抽取。

行车故障诊断领域文本与常规的识别语料不同，其中存在着实体嵌套、实体跨度大、关系重叠的情况，目前工业设备故障诊断领域的知识抽取方法难以对其进行有效知识抽取，因此本文以某企业行车故障诊断知识图谱构建的实际需求为引导，针对企业内部积累的行车故障调查单语料的实际特点，设计了知识抽取方法：

（1）在命名实体识别任务中，提出了融合强化学习的机器阅读理解模型，在段落级别的实体识别任务中，使用机器阅读理解的方式，对同一文本进行多次提问，将不同类型实体的识别进行隔离，解决多类型实体嵌套问题；通过指针网络解码进行实体头尾位置匹配，解决同类型实体嵌套以及长实体识别问题，并在模型中融入了基于策略梯度的强化学习方法，鼓励近似答案识别，进一步提升了模型识别效果；

（2）在关系抽取任务中，将关系抽取分为两阶段，先对句子中关系主体进行识别，再用机器阅读理解的方式把主体作为提示和文本内容一起输入到模型中进行提示学习来识别客体，将多实体对关系识别进行隔离，解决了句子级别的关系重叠问题；

（3）通过知识抽取得到行车故障诊断领域的实体和关系，将其整合成三元组，为行车故障诊断知识图谱构建提供数据支持。

1　语料特点

1.1　实体关系类型定义

行车故障调查单是表格类型的文档，每个文档记录了对应故障的详细信息，其中“故障名称”、“作业线”、“专业属性”、“设备停机时间”直接给出了对应值，可以作为结构化数据进行抽取，其余的 3 块内容“故障经过”、“原因分析”、“纠正措施”则是大段文字描述，内部都包含了多种类型的实体，因此将这 3 块内容作为 3 份数据集进行实体抽取，各数据集中实体类型见表 1。

表 1　行车故障诊断实体类型

Table 1　Type of crane fault diagnosis entities

数据集	实体类型
故障经过	故障行车、发生日期、开始时间、结束时间、故障细节、故障位置、故障表现、设备、部件、零件
原因分析	主要原因、次要原因、直接原因、间接原因、根本原因、管理原因、普通原因
纠正措施	解决措施、责任人

本文需要对行车故障调查单进行知识提取以构建故障诊断知识图谱，从知识图谱中获取故障发生的原因、解决措施等信息，辅助行车的故障诊断，针对该目标并结合故障调查单的实际内容格式，本文将实体间关系类型定义为如图 1 所示的关系类型。

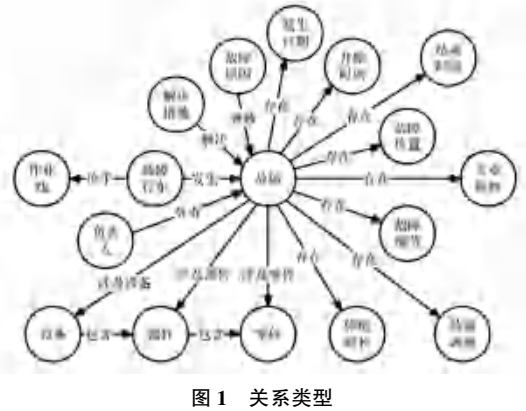

图 1　关系类型

Fig. 1　Type of relation

1.2　实体嵌套

行车故障诊断语料与常规实体识别语料不同，

在“故障经过”数据集的实体抽取上存在实体嵌套的情况,实体嵌套是指在一种类型的实体中还存在着同类或者其他类型的实体[12]。如图 2 所示,“故障细节”中嵌套着“故障位置”和“故障表现”,“故障位置”中又可能嵌套着“设备”、“部件”、“零件”这些实体,传统的序列标注识别方式往往是通过 CRF 层解码得到每个字符最有可能对应的标签,以这种方式识别,图 2 中“大车电气梁靠 3 号车方向车轮有异音”的“大”字会同时对应“故障细节”、“故障位置”、“设备”3 类实体的开头,而单个字符无法被同时解码为多个不同的标签,因此传统序列标注识别方法不适用于嵌套实体的识别。

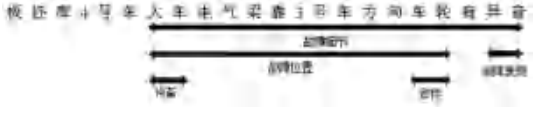

图 2 实体嵌套

Fig. 2 Nested entities

1.3 实体跨度大

行车故障诊断语料在“原因分析”和“纠正措施”数据集的实体识别上存在实体跨度大的情况,其中需要抽取的各类“故障原因”以及“解决措施”实体往往难以用几个字符概括,其实体长度一般从十几个到几十个字符不等,如果使用序列标注模型进行识别,由 CRF 层解码,可能会导致实体断裂,出现破键的情况,降低模型识别长实体的能力。

1.4 关系重叠

在行车故障诊断领域的关系抽取中,由于每份故障调查单都对应单次故障,因此除“设备”、“部件”、“零件”之间的“包含”关系,其他实体间关系在实体被抽取出来时就已经确定了,所以本文主要针对“包含”关系进行抽取,关系类别单一,只需找到“包含”关系的主体和客体即可,但其语料中存在着如图 3 所示的关系重叠问题,一个实体在两个关系中分别充当主体和客体以及一个主体可能会对应多个客体的情况,传统的关系抽取往往只关注于句子内单实体对之间的关系而无法处理多实体对的关系重叠问题[13]。

图 3 关系重叠

Fig. 3 Overlapping relations

2 行车故障诊断知识抽取方法

2.1 实体抽取方法

为解决命名实体识别任务中存在的实体嵌套和长实体识别问题,提出了 RoBERTa(Robustly Optimized BERT Pretraining Approach)-PN(Pointer Network)-RL(Reinforcement Learning)-MRC(Machine Reading Comprehension)模型,该模型将 RoBERTa 作为主干,将指针网络作为解码层,并融入了强化学习方法,模型整体结构如图 4 所示。

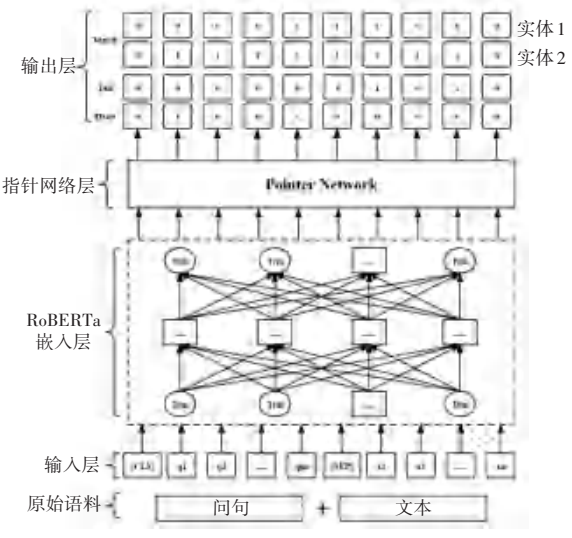

图 4 RoBERTa-PN-RL-MRC 模型结构

Fig. 4 Structure of RoBERTa-PN-RL-MRC

2.1.1 模型输入层

模型的输入是将问句 Q 和本文内容 X 拼接而成的字符串序列 $S = \{[\mathrm{CLS}], q_1, q_2, \cdots, q_m, [\mathrm{SEP}], x_1, x_2, \cdots, x_n\}$,其中[CLS]和[SEP]是 RoBERTa 模型中的特殊符号。每类实体对应了不同问句,通过将不同问句和文本内容一同输入模型进行多次提问,不仅在识别过程中增加了先验知识,而且把不同类型实体的识别进行了隔离,使不同类型实体的识别互不干扰,解决了多类型实体的嵌套问题。

2.1.2 RoBERTa 编码层

序列 S 进入模型中,首先需要经过 RoBERTa 编码层进行字嵌入编码,RoBERTa 是基于 BERT 的改进模型,相比于 BERT,其使用了更多的预训练语料和更长的训练时间,有很强的泛化能力,训练时去除了 NSP(Next Sentence Prediction)任务,增强了模型在单句上下文的建模能力,并且引入动态掩码机制,使得模型更加适应不同类型的任务,适合在小样本数据集上进行机器阅读理解[14]。输入序列 S 经

RoBERTa 编码层进行特征提取后，将输出中的问句内容以及两个特殊符号的嵌入表示去除就得到了文本的上下文表示矩阵 $\boldsymbol{E} \in R^{n*d}$，其中 n 为文本内容 X 的长度，d 为 RoBERTa 最后一层的向量维度。

2.1.3　指针网络解码层

将上下文表示矩阵 $\boldsymbol{E}$ 输入到指针网络层中进行解码。指针网络由两个线性层以及一个首尾匹配分类器组成，其需要预测文本中实体首尾边界，相比于预测整个序列标签的实体识别方法，其只需要学习如何指向序列中实体首尾位置，而不必学习如何生成完整的输出序列，因此模型参数更少，训练速度更快，能更好的捕捉到实体的范围和边界，有利于长实体的识别[15]。

在指针网络的解码过程中，以预测实体开始位置为例，首先要根据文本表示矩阵 $\boldsymbol{E}$ 得到每个位置作为开始位置的概率 P_{start}，计算方式如下：

$$P_{\text{start}} = \text{Softmax}_{\text{each row}}(\boldsymbol{E} \cdot \boldsymbol{T}_{\text{start}}) \in \mathbb{R}^{n*2} \quad (1)$$

其中，$\boldsymbol{T}_{\text{start}}$ 是一个 $d \times 2$ 的矩阵，内部权重由模型训练所得，P_{start} 每一行内容即为文本中每个位置能否成为实体开始位置的概率分布，对于 P_{start} 每一行都使用 argmax 函数计算，即可得到实体开始位置的预测序列 I_{start}，计算方式如下：

$$I_{\text{start}} = \{i \mid \text{argmax}(P_{\text{start}}^{(i)}) = 1, i = 1, \cdots, n\} \quad (2)$$

I_{end} 的预测方式与 I_{start} 类似，得到实体开始位置和结束位置的预测序列后，还需要对首尾位置进行匹配，由于本模型预测过程中可能存在同类型实体嵌套的情况，因此传统的就近匹配方法并不适用，对于每个位置的 $i \in I_{\text{start}}$ 和 $j \in I_{\text{end}}$，需要训练一个二分类器来预测匹配的概率，计算方式如下：

$$P_{i,j} = sigmoid(m \cdot concat(E_i, E_j)) \quad (3)$$

其中，$m \in \mathbb{R}^{1*2d}$，内部权重由模型训练所得，E_i 和 E_j 为文本中第 i 和 j 位置的字嵌入表示，通过这样的方式，开始位置和结束位置的匹配不再局限于一对一，其匹配方式变得十分灵活，可以解决同类型实体嵌套问题。

模型在进行训练时，主要预测的是实体的开始位置、结束位置以及首尾位置的匹配，因此损失函数定义如下：

$$L_{\text{total}} = \beta_1 L_{\text{start}} + \beta_2 L_{\text{end}} + \beta_3 L_{\text{span}} \quad (4)$$

其中，每项损失 L 均为预测结果和真实标签的交叉熵损失，β_1、β_2、β_3 作为超参数进行调整。

2.1.4　基于策略梯度的强化学习方法

初步模型的损失值由预测结果和真实标签之间的交叉熵损失来计算，但当模型预测结果与真实标签内容上相似，而首尾位置预测有偏差时会存在问题，因为模型对近似答案的惩罚和完全错误答案的惩罚相同。如在“41 号行车驾驶室联结螺栓盖开裂”中识别“零件”实体，正确答案是“联结螺栓盖”，如果模型预测“螺栓”，这是可以接受的答案，但对于交叉熵损失，模型对近似答案“螺栓”和错误答案“行车”的惩罚是相同的，这是不合理的。为鼓励模型对于近似实体的识别，本文使用基于策略梯度的强化学习方法，这种方法是对策略进行建模，然后通过梯度上升来更新策略网络的参数[16]；将强化学习损失融入到整体损失函数中作为辅助任务来对模型做进一步微调。

奖励在强化学习中用于评估一个动作的好坏程度[17]，这里将奖励定义为模型的预测结果和真实标签重叠部分的 $F1$ 值，具体计算方式如图 5 所示。

图 5　强化学习奖励计算方法

Fig. 5　Calculation method of reinforcement learning reward

强化学习的目标就是累计更多的奖励[18]，这就需要找到一个最优的策略 π_θ 使模型在不同输入状态 s 下输出的动作 a 所获得奖励 $R(a \mid s)$ 的期望最大，其中 π_θ 实质上是动作 a 的概率密度函数，在基于策略梯度的强化学习方法中，将神经网络近似成策略函数 π_θ，通过对目标函数求导得到策略梯度来优化网络参数，从而找到最优的策略函数 π_θ，本文以 RoBERTa 层来近似策略函数 π_θ，θ 是网络的参数集，策略梯度计算方法如下：

$$\nabla_\theta L_{rl} = -E[\nabla_\theta \log \pi_\theta(a \mid s) \cdot R(a \mid s)] \quad (5)$$

在实际计算过程中，式(5) 中的期望 $E[\ldots]$ 往往是难以计算的，常用的方法是使用蒙特卡洛近似，这种方法是通过从一个概率分布中随机抽取样本，利用这些样本的统计特性来估计所需要的量[19]。因此本文引入这种思想，在编码层的输出结果上随机采样多个动作来近似整体期望，以经验平均代替期望，并且根据 Greensmith 等[20] 的研究可知，在随机采样所得奖励的基础上减去一个基线，可以在不影响策略梯度结果的情况下有效减少策略梯度的方差，加快模型收敛。因此本文选择贪婪采样下预测结果的奖励值作为基线，不仅用以减小方差，而且能作为标准评判随机采样结果。

以预测实体开始位置为例，贪婪采样方法如式(2)所示，随机采样方法如下：

$$I_{\text{start\_random}} = \{i \mid \text{random}(P_{\text{start}}^{(i)}) = 1, i = 1, \cdots, n\} \tag{6}$$

其中,random(…)表示根据每个位置的概率分布进行一次随机抽样,将得到的 $I_{\text{start\_random}}$ 和 $I_{\text{end\_random}}$ 再经过式(3)进行匹配后就可得到随机采样的预测结果,贪婪采样的预测结果同理可得。最终经过蒙特卡洛近似并减去基线奖励后,策略梯度的计算方法如下:

$$\begin{aligned}\nabla_\theta L_{rl} = & -E[\nabla_\theta \log \pi_\theta(a|s) \cdot R(a|s)] \approx \\ & -\nabla_\theta \frac{1}{n}\sum_{1}^{n}[\log P_{\text{start}}(\theta) + \log P_{\text{end}}(\theta)] \cdot \\ & [R_{\text{random}}(a_i|s) - R_{\text{greedy}}(a_i|s)]\end{aligned} \tag{7}$$

其中,n 为随机采样的次数,$P_{\text{start}}(\theta)$ 和 $P_{\text{end}}(\theta)$ 分别是模型输出的头尾位置的概率分布,$R_{\text{random}}(a_i|s)$ 和 $R_{\text{greedy}}(a_i|s)$ 分别是第 i 次随机采样和贪婪采样下动作的奖励值。

通过把强化学习方法作为辅助任务来对模型进行微调,将强化学习的策略梯度作为模型损失的一部分,模型最终的损失函数如下:

$$L_{\text{total}} = \beta_1 L_{\text{start}} + \beta_2 L_{\text{end}} + \beta_3 L_{\text{span}} + \beta_4 L_{rl} \tag{8}$$

其中,β_4 也作为超参数进行调整。

2.2 关系抽取方法

由于每份行车故障调查单对应单次故障的详细信息,因此非“包含”关系可直接基于规则定义,而“包含”关系,虽然其关系类型单一,但存在着关系重叠问题,因此本文将关系抽取分为先识别主体再识别客体的两阶段,所使用的模型结构与命名实体识别任务中的模型相同,以 RoBERTa 为主干进行编码,以指针网络为解码层进行输出,在命名实体识别模型抽取结果基础上进一步标注“设备”、“部件”、“零件”之间的“包含”关系进行训练。先用模型进行主体识别,可能会识别出一个或多个主体,再将每个得到的主体分别和文本内容进行拼接输入到客体识别模型中,同样每次也可能会识别出一个或多个客体,通过这样的方式将不同实体对之间的关系抽取进行隔离,解决关系重叠问题,具体的抽取流程如图 6 所示。

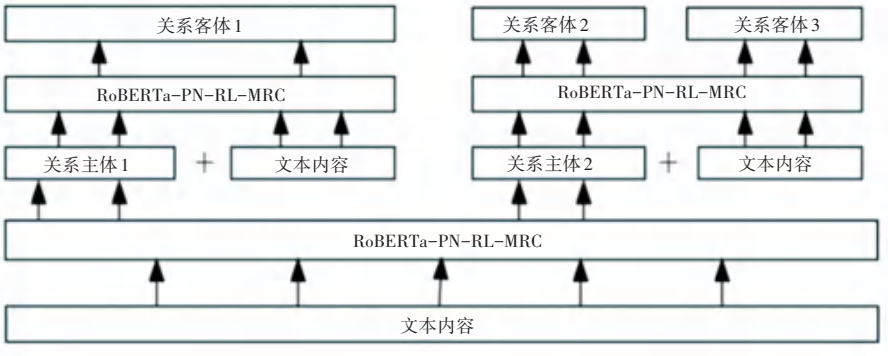

图 6 关系抽取流程

Fig. 6 Process of relational extraction

3 实验结果及分析

3.1 实验环境与评价指标

实验平台操作系统为 Ubuntu 18.04.4 LTS,支持软件版本为 Python 3.6、Pytorch 1.7.1。模型训练所用服务器的硬件配置 CPU 为 Intel(R) Xeon(R) CPU E5-2678 v3,内存 64 G,GPU 为 NVIDIA GeForce RTX2080Ti。

本文对模型性能进行评估的评价指标是准确率 P、召回率 R 以及综合考虑前两者指标的 $F1$ 值,计算方法如下:

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{11}$$

其中,TP 表示模型识别为正确的正样本数;FP 表示模型识别为正确的负样本数;FN 表示模型识别为错误的正样本数。

3.2 数据集

3.2.1 命名实体识别数据集

在故障经过、原因分析、纠正措施中分别抽取不同的实体类型,其中故障经过为嵌套实体数据集,数据数量为 1 020 条,原因分析和纠正措施是含长实体较多的扁平实体数据集,数据数量分别为 1 470 条和 480 条,以故障经过数据集为例,各实体类型对

应的问句见表 2。

表 2　故障经过数据集

Table 2　Data set of fault process

实体类型	问句
故障行车	实际生产中使用的带有特定编号的起重机
发生日期	行车发生故障的确定的日子、时期
开始时间	行车开始发生故障的时间点
结束时间	行车故障被解决的时间点
故障细节	描述故障包含设备的位置和发生的现象
故障位置	对行车发生故障的位置的整体描述
故障表现	行车故障的具体类型包括断裂、异音等
设备	行车的主要组成部分包括大车、小车、起升等
部件	行车中由若干零件装配在一起所组成的部件
零件	行车中最小的组成要素，不可拆分的单个质件

3.2.2　关系抽取数据集

关系抽取任务中，需要抽取“设备”、“部件”、“零件”之间的“包含”关系，因此先将所有文本语料都进行分句，再根据命名实体识别任务得到的结果对句子进行筛选，选择至少包含两个“设备”、“部件”、“零件”实体的句子，最后进行关系标注得到标注数据 231 条。

3.3　模型训练结果与分析

3.3.1　命名实体识别任务

模型训练将数据集按比例 8∶2 划分为训练集和验证集。经过多次实验，模型的主要训练参数配置见表 3。

表 3　命名实体识别参数配置

Table 3　Named entity recognition parameter setting

实验参数	参数值
学习率	8×10^{-6}
批次大小	1
最大长度	256
优化器	AdamW
失活率	0.2
迭代次数	20
损失函数比例	1、1、0.1、0.1

本文自建的 3 份数据集分别为故障经过、原因分析、纠正措施。故障经过作为嵌套实体数据集，选择文献[5]中的 Second-best Path 模型作为基准模型，并同时比较了强化学习损失对模型性能的影响；原因分析和纠正措施是含长实体较多的扁平实体数据集，因此把传统的 BERT-BiLSTM-CRF 序列标注模型作为基准模型，并比较了强化学习的效果，实验结果见表 4。

表 4　命名实体识别实验结果

Table 4　Experiment results of named entity recognition

故障经过			
模型	*P*/%	*R*/%	*F*1/%
Second-best Path[11]	75.97	63.64	69.26
RoBERTa-PN-MRC	91.41	85.14	88.17
RoBERTa-PN-RL-MRC	**92.59**	**85.71**	**89.02**
原因分析			
模型	*P*/%	*R*/%	*F*1/%
BERT-BiLSTM-CRF	73.91	80.95	77.27
RoBERTa-PN-MRC	84.91	**88.24**	86.54
RoBERTa-PN-RL-MRC	**93.62**	86.27	**89.80**
纠正措施			
模型	*P*/%	*R*/%	*F*1/%
BERT-BiLSTM-CRF	75.00	80.33	77.57
RoBERTa-PN-MRC	**79.90**	83.25	81.54
RoBERTa-PN-RL-MRC	79.05	**86.91**	**82.79**

由实验结果可知，本文所使用的模型不管在嵌套实体数据集上还是扁平实体数据集上均有良好的性能表现。在嵌套实体数据集上，本文模型在准确率、召回率、*F*1 值上均优于基线模型，其中 *F*1 值相较于基线模型有 19.76% 的提升；在两份扁平实体数据集上，本文模型性能同样均优于传统序列标注模型，在原因分析数据集上性能提升较明显，*F*1 值有 12.53% 的提升。

此外，实验结果还可证明强化学习方法的融入有助于模型性能的提升，在 3 份不同数据集上，融合了强化学习损失的模型在 *F*1 值上均优于不带强化学习损失的模型，在原因分析数据集上的效果最为明显，*F*1 值有 3.26% 的提升。上述结果证明强化学习方法的有效性，体现了其在机器阅读理解任务中作为辅助任务微调模型的价值，并最终抽取实体 2 564 个，属性 8 415 个。

3.3.2　关系抽取任务

将数据集按 8∶2 分为训练集和验证集，通过实验，模型训练的学习率为 3×10^{-5}，批次大小为 4，序列最大长度为 128，其余与命名实体识别任务相同。实验表明，先抽取关系主体再抽取关系客体模型整体的准确率为 91.11%，召回率为 95.35%，*F*1 值为 93.18%，基本能够满足行车故障诊断领域中“设备”、“部件”、“零件”之间“包含”关系的抽取，可在保证抽取效果的情况下解决关系重叠问题，并结合规则抽取出的关系，最终抽取关系 4 370 条。

4 结束语

为充分发挥行车故障调查单历史信息价值,本文对其进行知识抽取,以构建故障诊断知识图谱,方便知识检索及综合利用。重点针对语料中存在的实体嵌套和长实体识别问题,提出了融合强化学习的机器阅读理解模型 RoBERTa-PN-RL-MRC,能有效针对语料特点进行实体抽取;对于语料中存在的关系重叠问题,将关系抽取分为先识别主体再识别客体的两阶段,以解决多实体对之间的关系重叠问题。实验结果表明,本文所提出的模型在行车故障诊断领域能进行有效的知识抽取,并最终从 532 份行车故障调查单中抽取出三元组 12 785 个,能支持行车故障诊断知识图谱的构建。在未来的工作中,将继续扩充数据集语料,更加充分训练模型,并尝试研究不同问句构建方法对机器阅读理解模型性能的影响,以进一步提升模型识别效果。

参考文献

[1] 马亮, 彭开香, 董洁. 工业过程故障根源诊断与传播路径识别技术综述[J]. 自动化学报, 2022, 48(7): 1650-1663.

[2] 盛林, 马波, 张杨. 基于知识图谱的旋转机械故障诊断方法[J]. 机电工程, 2022, 39(9): 1194-1202.

[3] GAO W, ZHENG X, ZHAO S. Named entity recognition method of Chinese EMR based on BERT - BiLSTM - CRF [C]// Proceedings of Journal of Physics: Conference Series. IEEE, 2021, 1848(1): 012083.

[4] 祁鹏年, 廖雨伦, 覃飙. 基于深度学习的中文命名实体识别研究综述[J]. 小型微型计算机系统, 2023, 44(9): 1857-1868.

[5] 余诗媛, 郭淑明, 黄瑞阳, 等. 嵌套命名实体识别研究进展[J]. 计算机科学, 2021, 48(S2): 1-10,29.

[6] JU M, MIWA M, ANANIADOU S. A neural layered model for nested named entity recognition [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). IEEE, 2018: 1446-1459.

[7] SHIBUYA T, HOVY E. Nested named entity recognition via second-best sequence learning and decoding [J]. Transactions of the Association for Computational Linguistics, 2020, 8: 605-620.

[8] WANG H, QIN K, ZAKARI R Y, et al. Deep neural network-based relation e-xtraction: an overview [J]. Neural Computing and Applications, 2022,34(6): 1-21.

[9] 苏杭, 胡亚豪, 谢艺菲, 等. 利用提示调优实现两阶段模型复用的关系实体抽取方法[J]. 计算机应用研究, 2022, 39(12): 3598-3604.

[10] 李福琳. 实体抽取及关系发现关键技术研究[J]. 信息技术与信息化, 2019 (10): 220-221,224.

[11] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme [J]. arXiv preprint arXiv: 1706. 05075, 2017.

[12] 闫璟辉, 宗成庆, 徐金安. 中文医疗文本中的嵌套实体识别方法[J]. 软件学报, 2024, 35(6): 2923-2935.

[13] 冯钧, 张涛, 杭婷婷. 重叠实体关系抽取综述[J]. 计算机工程与应用, 2022, 58(1): 1-11.

[14] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach [J]. arXiv preprint arXiv: 1907. 11692, 2019.

[15] VINYALS O, FORTUNATO M, JAITLY N. Pointer networks [J]. Advances in Neural Information Processing Systems, 2015, 2(28):2692-2700.

[16] KIM D K, LIU M, RIEMER M D, et al. A policy gradient algorithm for learning to learn in multiagent reinforcement learning [C]// Proceedings of International Conference on Machine Learning. IEEE,2021: 5541-5550.

[17] 李明阳,许可儿,宋志强,等. 多智能体强化学习算法研究综述[J]. 计算机科学与探索,2024,18(8):1979-1997.

[18] 李茹杨, 彭慧民, 李仁刚, 等. 强化学习算法与应用综述[J]. 计算机系统应用, 2020, 29(12): 13-25.

[19] MOHAMED S, ROSCA M, FIGURNOV M, et al. Monte carlo gradient estimation in machine learning [J]. Journal of Machine Learning Research, 2020, 21(1): 5183-5244.

[20] GREENSMITH E, BARTLETT P L, BAXTER J. Variance reduction techniques for gradient estimates in reinforcement learning [J]. Journal of Machine Learning Research, 2004, 5(9):1.