文章编号:1671-4598(2024)06-0027-08 DOI:10.16526/j.cnki.11-4762/tp.2024.06.005 中图分类号:TP391.1 文献标识码:A

基于模型的机床故障案例命名实体抽取方法比较研究

尹昱东[1]，王保建[1]，李珂嘉[1]，王紫平[1]，张小丽[2]
(1. 西安交通大学 机械工程学院，西安 710049；
2. 长安大学 道路施工技术与装备教育部重点实验室，西安 710064)

摘要：机床出现的故障大多有先例，但故障案例分散，不同工厂又不数据共享且没有标准的数据库管理，以至于对于已有的相似故障，工厂仍需要按照未知故障进行停机维修；因此，急需一套标准服务平台能够集合大量故障案例，同时实现更新维护，增添新故障，以供各工厂做故障参考，尽可能降低维修成本以及时间开销；通过将计算机领域较为流行的知识图谱运用到机床故障诊断领域，全面运用机床故障诊断案例知识，构建以故障现象、故障发生原因以及解决方案为核心的机床故障诊断网络，实现快速确认故障发生部位，提供合理的故障解决方案，提高制造业的生产效率；使用爬虫技术获取故障案例数据，采用 BIO 标注法完成样本标注，分别使用 Bilstrm-crf、Vgg16 以及 Bert 模型完成实体抽取任务，并对上述模型准确率从多个角度进行对比，将知识导入 Neo4J 图数据库并建立针对机床故障的知识图谱，最终实现知识图谱可视化。

关键词：机床故障；知识图谱；Bert 模型；Neo4j 图数据库；命名实体抽取

Comparative Study on Model-Based Named Entity Extraction Methods for Machine Tool Fault Cases

YIN Yudong[1], WANG Baojian[1], LI Kejia[1], WANG Ziping[1], ZHANG Xiaoli[2]
(1. School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
2. Key Laboratory of Road Construction Technology and Equipment of the Ministry of Education, Chang'an University, Xi'an 710064, China)

Abstract: Most machine tool failures have the characteristics of scattered fault cases, no data sharing in different factories, and no standard database management. Aimed at existing similar failures, it is still necessary to shut down for maintenance in factories by unknown failures. Therefore, it is urgent for a standard service platform to collect a large number of fault cases, update maintenance and add new faults at the same time, and provide a fault reference for each factory and reduce maintenance and time costs as much as possible. This paper applies the popular knowledge graphs in the computer field to the fault diagnosis networks of machine tools, comprehensively uses the case knowledge of machine tool fault diagnosis, and constructs a machine tool fault diagnosis network with fault phenomena, reasons and solutions as a core to quickly identify the fault location, provide reasonable fault solutions, and improve the production efficiency of manufacturing industry. The crawler technology is used to obtain the fault case data, the BIO tagging method is used to complete the sample tagging, the Bilstrm-crf, Vgg16 and Bert models are used to complete the entity extraction task, respectively, and the accuracy of the above models are compared from multiple aspects, the knowledge is imported into the Neo4J diagram database, it builds the knowledge graph for the machine tool fault, and finally realizes the visualization of knowledge graph.

Keywords: machine tool fault; knowledge graph; Bert model; Neo4J graph database; named entity extraction

0 引言

随着“中国制造 2025”的提出，传统制造领域开始实现数字化管理，并引入人工智能技术。在机床故障诊断领域，由于机床在运行时常出现故障，可是故障知识又过于分散，数据量大，缺乏管理，一旦出现故障，通常需要专家进行诊断，因此对诊断人员专业水平要求很高，但领域内专业人士稀少，直接影响到故障诊断效率，也导致机床维护成本居高不下。在当前的机床数字化管理领域，如何低成本维护，以及降低诊断人员的门槛成为数控机床故障领域智能化的关键之一。能够集合大量故障案例，同时可

收稿日期:2023-04-03； **修回日期**:2023-06-12。
基金项目:陕西省自然科学基础研究计划项目(2021M-169)；陕西省自然科学基础研究计划项目(2023-JC-YB-477)；2022 年西安交通大学本科实验实践与创新创业教育教学改革研究专项项目(22SJZX10)。
作者简介:尹昱东(1983-)，男，博士，工程师。
引用格式:尹昱东，王保建，李珂嘉，等.基于模型的机床故障案例命名实体抽取方法比较研究[J].计算机测量与控制，2024，32(6)：27-34.

以实现知识更新、知识推理，以供各工厂做出故障诊断参考的标准专家系统平台能够发挥重要作用。

对于数控机床设备的诊断维护有两种方法：人工诊断法和智能诊断法。人工诊断主要为事后诊断，即设备故障发生后进行诊断，有设备外观故障检查、软故障检查、数据检查等，极度依赖操作者专业知识和经验。设备外观检查包括观察设备外形、工件加工质量、操作者通过听觉和嗅觉以及功能程序测试等方式进行判断机床是否故障[1]。软故障检查指操作者通过翻阅历年故障案例数据集进一步判断机床故障。机床数据检查适用于带数控系统的机床，CNC 系统可以针对机床故障反馈一定的报警代码，通过代码进一步确认故障。由此可见，人工诊断效率并不高，近年来，不再成为机床诊断方式主流。

智能故障诊断通常包含 3 个环节：信号获取、特征提取、故障识别与预测[2]。信号获取通常采用获取机床振动信号包含位移、速度和加速度信息，通过信号反映设备状况；特征提取对象包含时域特征、频域特征以及时频域特征等；故障识别与预测是以提取的特征为基础，使用合适的人工智能技术，搭建诊断模型，进行故障识别和预测，尽可能在出现故障前对机床故障进行警告，操作人员可以根据警告决定机床工作时间，尽可能实现统一时间维修，减少维修次数以及维修成本。

智能诊断法成为主流后，陆续出现了各种各样的机床故障诊断方法。盛博等人[3]提出将图论引入故障诊断，基于图论建立多种描述故障模型，同时引入全局风险影响度概念，降低风险主观性因素影响，以及相比于传统方法，提高故障定位准确度。文献［4］使用循环神经网络技术，借助“门”判别结构同时联合模糊性特征空间，实现数控机床故障在线诊断。文献［5］采用 CBR 与 RBR 结合的推理机制，建立专家系统知识库。文献［6］将故障树与案例推理的方法引入故障诊断当中，实现规则与案例相结合的推理机制，开发了数控镗铣床的原型系统。

目前的推理仅实现单一推理或者单对多的推理，准确率并不高，且技术不够成熟，例如多对多的关系推理还没实现，描述故障事件之间的关系方法不够灵活，故障知识库还有大量关系没有挖掘；机床的故障诊断系统繁多，故障数据来源结构复杂，格式众多，且没有较为成型的标准数据库，不易于数字化管理。随着时间推移，机床的故障诊断系统也应该能够实现低成本自我更新，实现新故障知识融入和已淘汰的故障知识删除，大幅度降低知识库存储的内存容量，而不是不断提出新系统去取代旧系统。

人工智能领域存在大量先进技术，将这些技术合理应用到机床故障诊断领域当中，提高故障识别准确率，挖掘潜在故障特征以及关系，也是实现机床故障诊断智能化的一个趋势。例如将知识图谱技术运用到机床故障诊断领域，可以全面运用机床故障诊断案例知识，构建以故障现象、故障发生原因以及解决方案为核心的机床故障诊断网络，可以实现快速确认故障发生部位，提供合理的故障解决方案，最终提高制造业的生产效率。

知识图谱概念由 Google 于 2012 年提出[7]，通过图的结构，从数据当中提取实体、关系等信息，并建立关系模型，实现互联网知识的有效利用。常见的通用知识图谱有 Freebase、Wikidata、DBpedia、YAGO 等[8-9]。行业知识图谱有 IMDB、MusicBrainz、ConceptNet。

知识图谱的构建技术主要在知识的抽取、表示、融合以及推理这 4 个环节，其中知识抽取主要指的是从开放数据领域当中自动抽取所需要的知识，其中又包括实体抽取、关系抽取以及属性抽取等关键环节。

实体抽取主要使用命名实体识别领域的技术。命名实体识别是从大量语料当中寻找代表对象的指示词。文献［10］将实体抽取的方法划分为基于规则的方法和基于统计机器学习的方法。基于规则的方法往往准确率会更高，但是需要大量人力且覆盖范围有限，而基于统计机器学习根据样本量的多少，可以划分为常规深度学习方法以及少数样本学习方法甚至无样本学习。文献［11］基于规则的方法首次设计了一款自动挖掘文本当中公司名的实体抽取系统，使用了自启发式算法。关于统计机器学习的方法，最近有清华大学开发的 ERNIE 模型，将知识图谱与 Bert 模型结合诞生出 K-Bert 模型，集合知识图谱对知识的理解优势和 Bert 模型对字的优秀推理能力，实现多词推理，对语料库进行知识抽取，且准确率高[12]。

关系抽取在实体抽取之后，继续从原来的语料挖掘已抽取的实体之间的关系信息。一个完整的知识单元必须包含实体以及关系，关系既可以是实体与实体之间的关系，也可以是连接实体与特征的桥梁。关系抽取研究方法早期采用人工提取语料的关系，要求专业人员必须对特定领域语言有深刻理解，且专业表达语言能力丰富。之后，统计学习机器模型替代了人工成为关系抽取的主流。自然语言技术引入关系抽取领域，随之诞生许多机器模型。这一类机器模型有马尔可夫逻辑网，无监督学习模型 StatSnowball，EntSum 模型以及 Markov 逻辑 TML 模型。

属性抽取也是在实体抽取之后，从原先的语料集挖掘属于实体的属性信息，比如针对某个旅游景点，可以从开放知识库中得到开放时间、所属地名、旅游景点评级、相邻著名建筑、文化背景等信息，通过收集这一类信息，完成对实体的完整性评述。属性抽取可以等价于关系抽取，将属性值转化为第二实体，而实体与属性值之间的联系理解为两个实体之间的关系。比如，文献［13］使用条件随机场与支持向量机模型结合，将实体、属性与属性值视为 3 类实体，实现旅游景点属性抽取转化为实体关系抽取，依靠支持向量机模型做实体关系预测，最终实现 84.4％的分类精确度。

知识图谱的知识抽取环节目前有很多先进技术，不断将自然语言技术与深度学习技术作为知识图谱技术支撑，提高知识图谱构建的精确度。另外，知识图谱对知识具备优秀的理解能力，诞生出以知识图谱作为自然语言技术支

撑的新研究路径。知识图谱开始应用于搜索引擎，后来逐渐参与智能问答系统，最近开始流行于推荐系统、辅助决策等领域。知识图谱不仅适用于通用科普领域，还适用于专业领域[14]。文献［15］根据半结构化装配数据进行知识抽取，实现了从三维模型到知识图谱的构建。

本文以机床故障案例为研究对象，利用知识图谱可视化优点和图数据库的高查找效率，完成机床故障诊断知识图谱构建，构建过程涉及本体建模、数据获取、数据处理、知识提取、知识存储、可视化表达等重要环节，通过图的结构，能够有效、灵活表达零件之间、零件与故障之间、故障与解决措施之间等关联关系，挖掘知识图谱潜在关系，有效扩展知识图谱，最终依靠故障诊断知识图谱有效诊断当前故障现象并给出合理的故障维修方案。

在构建知识图谱过程中，数据来源十分重要，通过收集关于故障案例及专业人士经验知识的半结构化数据和无结构化数据，搭建数据库，实现对数据进行知识提取，使用 Bert 模型进行命名实体提取，另外采用了多种模型包括 Bilstrm-crf 和 Vgg16 模型进行测试分析，比较模型的准确率。

1　数据预处理与知识图谱构建

知识图谱的数据获取的类型可以分为 3 种，分别是结构化数据、半结构化数据以及非结构化数据。非结构化数据大部分指的是以文本形式存在的数据，常见的有互联网上的文章，而结构化数据有很明显的数据结构，主要指的是数据库当中存储的数据，由于企业或者研究机构要求不同，因此结构化数据结构类型很多，因此如何将不同数据结构同一化成为一个关键，即异源融合。而半结构化数据可以理解为有结构的文本，是介于结构化数据和非结构化数据两者之间，结构化数据和非结构化数据同时存在于同一数据当中。目前关于机床故障诊断的结构化数据很少，没有标准成型的相关数据库，公开的数据库更是少之又少，机床故障诊断常出现在实际工厂加工期间，也就是说大多的机床故障诊断案例以书籍、经验贴以及工厂维修表等文本形式存在。因此为了搭建基于机床故障诊断专业领域的知识图谱，本研究选择的数据来源为半结构化数据和非结构化数据，设计了一种机床故障诊断知识图谱构建的方法，实现从初始数据当中提取知识图谱所需要的存储单元—三元组，并将三元组存入图数据库，完成知识图谱的搭建。

1.1　机床故障知识图谱构建流程

知识图谱的搭建方式可以分为自顶向下和自底向上。自顶向上的构建方式，就是从网站获取数据，从数据中提取知识，存入知识库。而自底向上的构建方式，需要以其他成型的数据库为基础，主要采用数据迁移的方式，将需要的数据存入知识库当中。前面谈到，机床故障数据暂时没有公开以及认可度高的数据库，因此，本文采用自顶向上的方式搭建知识图谱。

知识图谱构建包含数据获取、数据清洗及加工、知识抽取、知识存储以及可视化表达。文献［16］提出采用自底向上的方式构建知识图谱需要进行多次迭代，每次迭代都包含信息提取、知识融合以及知识加工共 3 个环节。构建知识图谱包含知识获取、知识表示、知识融合、本体构建、知识推理、知识更新以及质量评估等 7 个方面。

参考以上几种思路，同时结合机床故障诊断问题的多种需求，本文设计了较为合理的构建数控机床故障诊断知识图谱的流程，即本体建模、数据获取、数据处理、知识提取、知识存储、可视化表达。具体流程如图 1 所示。

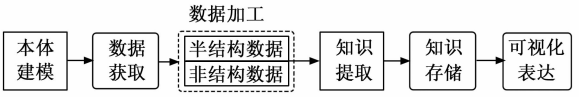

图 1　机床故障诊断知识图谱搭建流程图

本体建模根据项目图谱设计需求构建知识图谱的模式层；数据获取包括从专业社区获取机床故障相关数据以及从专业书籍获取机床故障案例集；数据处理指的是对收集到的数据进行预处理，包含去除停用词和无意义符号，文本分词处理以及样本标注等；知识提取指的是从数据集当中提取构建知识图谱所需要的实体信息和关系信息；知识存储将知识提取得到的信息转化为三元组的形式，存入搭建好的数据库当中；图谱可视化表达，将存储好的知识进一步以图的形式展示，实现机床故障诊断知识体系的可视化表达。

1.2　机床故障知识本体建模

知识图谱可以划分为模式层和数据层，模式层最为重要，作为知识图谱的核心，是数据层上进一步提取形成的知识，常用本体库来表示。相比较于数据层上的知识库，本体库中的知识更加规范，更有利于知识推理以及实体规则约束，同时其冗余知识更少。

采用网络本体语言 OWL 描述机床故障的本体，创建本体模型，以方便之后的故障知识领域的确定和故障实体及关系知识抽取。根据机床故障案例，以机床故障为顶级本体，另外提取机床故障现象、故障零部件、故障原因、故障操作、故障维修的核心信息，使用容易上手操作的本体开发工具 Protégé 构建本体模型。将故障信息分为故障现象、故障原因、故障维修、故障操作和机床故障零部件 5 个环节，通过 rdfs：subClassOf 将这些环节与机床故障联系。

核心本体类的具体含义如下：

1）故障现象指机床出现故障时警告代码、机床本身状况以及加工工件，比如机床停机、打刀、大批量工件不符合加工标准，故障现象会尽可能从多个方面描述机床故障，方便区分不同的故障以及提供合适的故障参考；

2）故障设备指机床类型、机床型号以及出现故障的零部件；

3）故障原因指产生机床故障的原因，大多为专家针对故障现象提供的解释；

4）故障操作指机床发生故障前对机床进行的操作，便于后续对故障维修提供科学的参考依据；

5）故障维修指让机床正常运行或工件加工正常所需的具体操作。

Protégé 建立的本体模型如图 2 和图 3 所示。

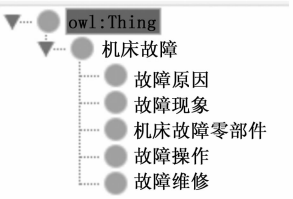

图 2 机床故障特征类

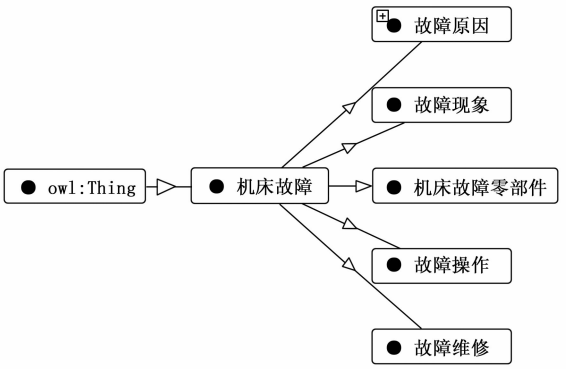

图 3 机床故障本体模型

1.3 数据获取

半结构化数据主要在文本数据当中又存在一定的分类，例如本文所用半结构化数据选自《西门子数控机床故障诊断 385 例》，在该书中所提供的故障案例都会包含故障现象、机床名称、数控系统、故障处理以及实际诊断过程等内容，因此只需要通过文本内容识别方法，识别到相应的标签，将标记好的文本内容存储到事先布置好的数据库当中。采用 MySQL 数据库进行存储，将表定义了 4 个列，分别为机床名称、故障现象、故障处理以及案例 id，将 id 设置为主键用来区分不同案例。

非结构化数据大量分布在网络文本上，数据的结构极其不规则，分布不集中，且有效信息少，因此采用爬虫技术，实现从指定范围内的网页中抓取指定的信息。为了抓取关于机床故障的非结构信息，本文采用 Scrapy 爬虫应用程序框架。Scrapy 基于 Python 语言开发的能够快速实现 web 抓取，主要用于抓取 web 站点以及相关页面的数据，只需少量的规范代码，就能够快速抓取信息。

Scrapy 五大基本组成模块有调度器、下载器、爬虫 Spider、实体管道、Scrapy 引擎。调度器用于决定需要抓取的网址顺序，根据分配好的顺序，向相关网址发送请求。下载器顾名思义就是下载网络资源，将数据下载到管道。爬虫 Spider 面向用户，由用户决定从特定网页中抓取特定的资源，同时也可以提取链接，同时可以实现基于广度优先或者深度优先等策略。实体管道用于处理爬虫提取的实体数据，主要有持久化实体、验证实体的有效性、清除不需要的信息等。

为了实现机床故障数据的高效获取，获取高质量文本，本文的非结构化数据主要来源于专业知识网站，如百度知道、贴吧、知乎等问答社区，主要采用广度优先探索策略，具体使用到的部分社区 URL 如表 1 所示。

表 1 部分社区 URL 表

数据来源	URL
知乎	https://www.zhihu.com/
百度贴吧	https://tieba.baidu.com/
百度知道	https://zhidao.baidu.com/
CSDN	https://so.csdn.net/

1.4 定义实体与标注模式

在对机床故障数据进行标注以及命名实体识别前，需要定义实体的类别以及标注的模式。本文将实体类别分为机床设备、故障现象以及故障操作三大类，分别定义为 Dev，Phe 以及 Ope 三类名词。

另外采用 BIO 标注对实体进行标注[17]。序列标注本质是给序列中每一字进行分类，即逐帧分类。训练一个判别器，输入一个字，通过结合该字的上下文，最终给出该字的类别。B、I、O 三大标签来区分实体首部、实体非首部以及非实体部分，需要在标签后添加相应的实体名词来区分三大实体类别。此外用 Part 表示机床设备类型实体，Phe 表示机床故障现象类型实体，Ope 表示故障相关操作类型实体。综上所述，可以得到大小为 7 的标签集合（B-Part、I-Part、B-Phe、I-Phe、B-Ope、I-Ope、O）。

通过爬虫技术收集到的非结构化数据以及《西门子数控机床故障诊断 385 例》抽取得到的半结构化数据中存在大量的噪音，口语类型数据较多，需要数据清洗。清洗主要有处理字母大小写，删除无意义的符号；随后对数据进行分词处理，使用 jieba 模型进行分词，但 jieba 模型对于机床领域并不完全适用，例如“加工中心”在 jieba 模型当中会分为“加工”和“中心”两个名词，因此在分词后需要对数据信息人工修改；接下来实体识别采用的核心思想是对模型进行监督训练，用部分的半结构化数据对模型进行训练和测试，之后用训练好的模型对其余半结构化数据和非结构化数据提取实体信息。因此，需要对训练样本和测试样本进行人工序列标注。以句子“行程开关经过一段时间使用后，紧固螺钉松动”为例，标注结果如表 2 所示。

表 2 序列 BIO 标注示例表

标注	句子	标注	句子	标注	句子	标注	句子
行	B-Part	过	O	使	O	螺	I-Part
程	I-Part	一	O	用	O	钉	I-Part
开	I-Part	段	O	后	O	松	B-Phe
关	I-Part	时	O	紧	B-Part	动	I-Phe
经	O	间	O	固	I-Part		

2 基于模型的命名实体抽取测试

命名实体识别，又称实体抽取，属于自然语言处理当中的一项不可或缺的任务，究其本质就是从文本语料当中

提取所需要的特定实体，通常包括人物、特殊物种、机构、设备等类型。而本文需要使用实体抽取从文本提取关于机床故障信息，得到面向机床故障专业领域的知识图谱所需要的实体信息。分别使用 Bilstrm-crf、Vgg16 以及 Bert 模型完成实体抽取任务

2.1　BiLSTM-CRF 模型

BiLSTM-CRF 模型本质上就是在双向 LSTM 模型上加一层 CRF 模型[18]，具体架构如图 4 所示。

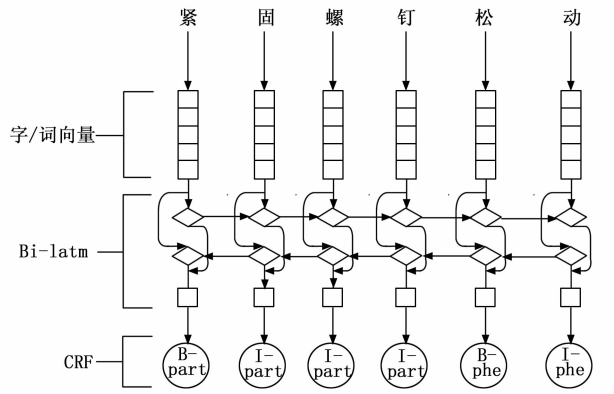

图 4　BiLSTM-CRF 架构

首先在输入层上通过 Embedding 层，将中英文字词投射到向量空间，即转变成计算机可以认识的符号，实现对文本的数字表示，常用的使用 Word2vec 和 One-hot 的形式来表示向量。其中，One-hot 又被称为 1-of-N representation，依照字典长度 n，给文本当中每个字词建立一个长度为 n 的向量，向量中为 1 的位置便是该字词在字典当中的位置，其余为 0。按着这种形式，依次替换句子当中的字，就可以将句子投射到向量空间。虽然这种方式表示词向量非常简单，但是当字典容量很大，就会导致所需要的内存庞大，例如当字典容量达到百万甚至千万级别，每个字词需要用相应的数量级别长度的向量表示，而且向量除了 1 就是 0，效率不高，会导致在神经网络中难以收敛。Word2vec 恰好解决了 One-hot 的缺陷，通过将每个词映射到短词向量上，依次重复上述操作，最终将生成的词向量构成向量空间，进一步依据统计学的理论来研究字词之间的相互关系。

在 Embedding 层后，前向 LSTM 和后向 LSTM 构成 BiLSTM，正因为前向和后向的存在很好地解决了自然语言任务处理当中，只能提取单项信息的难题。LSTM 模型可以理解为特殊的 RNN 模型，一般常见的 RNN 模型会受到短期记忆影响，将前一次结果带入下一次训练当中，能够存储一定的信息，实现“追根溯源”的功能，但是往往因梯度消失，可利用的历史数据有限。而 LSTM 模型可以记住长期信息，比 RNN 多了一个细胞状态，细胞状态的更新缓慢，记载之前状态的信息。另外，由于神经网络非线性拟合能力突出，BiLSTM 可以将前面得到的向量复杂度提升，精度提高，得到一个高维度的向量空间。但是存在一些问题，例如前面标注使用到的 BIO 方法，使用 BiLSTM，训练得到的结果会出现 B 标注之后紧跟着 B 的情况，实际上，B 作为实体的首字，其之后只可能为标注 I；还会出现 O 标注之后紧跟着 I 的情况，而实际上，I 应该在 B 之后或者在 I 之后。因此，需要在 BiLSTM 后面添加 CRF 模型，CRF 模型当中存在的特征函数，可以很好地限制不同字的标签之间的关系。因此，结合两者，可以集合双方，提高模型训练的效率。

2.2　Vgg16 模型

VGGNet 一共有 5 种神经结构网络，分别对应 A-E，就包含 Vgg16 网络结构，其中 16 指的是神经网络的深度[19]。由于 VGGNet 使用小型卷积核，通过增加网络深度来提高模型性能，实现更多的非线性映射，因此目前正广泛应用于分类任务和定位任务。具体 Vgg16 模型的网络结构如图 5 所示。

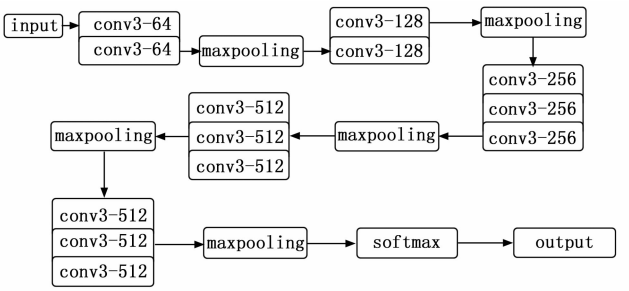

图 5　Vgg16 模型网络架构

Vgg16 的卷积层的卷积核大小都是 3×3 并具备 3 个通道，将几个相同的卷积层集合成单个模块，比如输入层之后的两层卷积层神经元都是 64 个卷积核，形成模块 1，每个模块之后都会带一个最大池化层，池化核大小为 2×2，最大池化层作用是对模型进行降维，降低计算机复杂度。Vgg16 有 13 个卷积层、3 个池化层，这 16 层主要做特征提取，后面添加全连接层，做模型分类。Vgg16 整体神经元数量在千万级别，因此模型具备良好的拟合能力，但也因此训练时间长，对存储容量要求高，一般设备仅能支撑单个模型运行，且需要进行模型限制分配内存，才能正常训练。

2.3　Bert 模型

Bert 模型具备 BiLSTM 和 Transformer 两个模型优点[20]，输入为每一个 token 的表征，其中 token 具备 3 个部分，其组成如图 6 所示。词嵌入（Token Embeddings）为系统根据当前字词在字典的位置，赋予的编号，主要用来区分不同的字词。E_机床代表机床这个词的编号。分割嵌入（Segment Embeddings），用来区分不同的句子，确定句子的句序，其中 E_A 代表第一个句子，E_B 代表第二个句子，E_A 在 E_B 之前。“机床启动，显示屏不显示”，“机床启动”在“显示屏不显示”前。实际使用中，分割嵌入使用率不高，效果不如其他两个嵌入。位置嵌入（Position Embeddings）记录当前词在该句子的位置，如“机床”在示例句子当中，占 1 号位置，在该环节每个字词的位置信息变化特征向量。SEP 用来拆分句子，CLS 表示句子的开始。通过 token 表征不难发现，对比于之前的 BiLSTM-CRF 模型，Bert 模型以字词为单位，正因如此，Bert 模型训练过程常被人称为模型在做“完形填空”题。

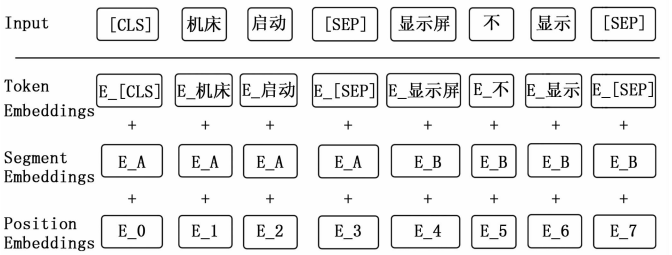

图 6 Bert 模型 token 表征组成

Bert 模型以编码器为基础，本质上是由多个编码器叠加而成，使用“注意力”模型做特征提取。该模型包含两个工作，分别是遮蔽语言（MLM）和下一个句子预测（NSP）。遮蔽语言算法思路是将 85%词嵌入不做处理，将剩余词嵌入分 3 种情况处理：

1）80%的词嵌入直接替换为［mask］；如“机床 启动”，处理之后，变为“［mask］启动”；

2）10%的词嵌入直接替换为其他任意单词；如“机床 启动”，处理之后，变为“电源 启动”；

3）10%的词嵌入保留 token，比如“机床 启动”，处理之后变为“机床 启动”。

这种处理方式，是为了减少模型出现没有见过单词的可能性。通过加入随机 token，是为了让编码器记住对每一输入词的表征。

下一个句子预测判断句子的序列，比如句子 A 是否是句子 B 的上文。训练时随机从语料库当中抽取连续的两句话，其中语料库经过预处理，经过处理后，50%的样本当中两句话符合上下文关系，剩余不符合。

Bert 模型存在一些缺点，不容易进行底层修改，常见用法是官方提供的 Bert 模型做微调，依靠相关领域的数据对已经通过大数据训练好的 Bert 模型做进一步训练，且收敛慢，需要高算力的设备支撑。

2.4 实验参数设定

命名实体抽取实验的所有模型都是基于 TensorFlow 进行构建，其中 Keras 版本为 2.3.1，TensorFlow 版本号为 2.2.0，keras _ contrib 版本号为 0.0.2。

实验选用的模型有 Vgg16、BiLSTM-CRF、Bert 模型，为了避免或者减轻过拟合现象在模型训练带来的影响，每个模型都适当加入池化层，每个模型的具体参数如表 3 所示。模型统一使用 Adam 优化器和 categorical _ crossentropy 损失函数，Batch size 参数越大，对计算机的算力要求越高，同样词向量维度也是如此，经过多次测试，为了保证设备正常运行且保持模型高效率，Batch size 参数设置为 64，词向量维度设置为 32。另外序列长度代表每个输入样本语料长度，通常将长度低于设定值的语料进行一边补零处理，对长度高于设定值的语料删除多余参数，保证每个语料长度都为序列长度值。标签类别数对应着分类数，本章标签数为 7（2 * 3+1），其中 2 刚好对应 BIO 标注的 B 和 I，3 为实际类别，如机床设备、故障现象、故障原因，1 为 O，O 在 BIO 标注为其他实体或者非实体文字。

表 3 模型关键参数表

参数	描述	值
Epoch	总迭代次数	50
Dropout rate	Dropout 率或者丢失率	0.217
Adam value	Adam 优化器参数	1×10^{-4}
Batch size	每批次训练样本大小	64
Num classes	标签类别数	7
Kernel size	卷积核尺寸大小(默认长和宽一样)	5
Seq length	序列长度	100
Embedding dim	词向量维度	32
Vocab size	字典大小	5 000
Learning rate	学习率	1×10^{-3}

3 实验结果与分析

将 Vgg16、BiLSTM-CRF 以及 Bert 这 3 个模型进行对比实验，以机床故障诊断案例为数据集，进行实体关系抽取，使用测试集测试结果如表 4 所示。

表 4 不同模型实体提取结果比较

模型	准确率/%	损失
Vgg16	92.06	0.339 9
BiLSTM-CRF	92.59	0.336 9
Bert	95.24	0.193 8

Bert 模型相比于其他方法，具备更高的准确率，Vgg16 与 BiLSTM-CRF 模型性能接近，准确率都低于 Bert 模型，但是实际 Vgg16 训练所需的时间远超 BiLSTM-CRF 模型。Bert、BiLSTM-CRF 以及 Vgg16 这 3 个模型在训练过程中准确率和损失值的变化如图 7 和图 8 所示。Bert 模型训练过程准确率曲线平滑，且一直高于其他两个模型，而 Vgg16 模型在训练初始阶段，准确率几乎不变，出现收敛慢的现象，从第五次迭代开始，才迅速提高准确率。BiLSTM-CRF 模型曲线在开始两次的训练当中，也出现停滞现象，之后就保持稳定且缓慢地提升，在第五次迭代前，模型性能高于 Vgg16，之后略低于 Vgg16。

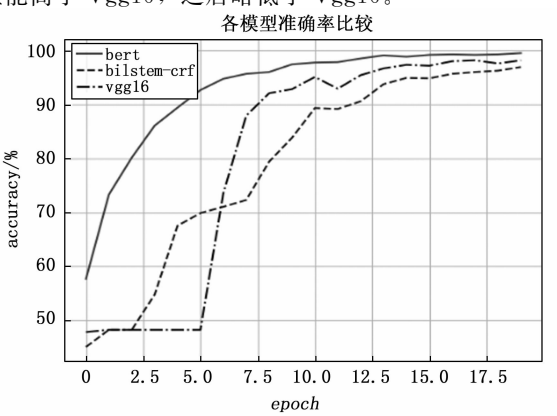

图 7 3 个模型训练过程准确率变化比较

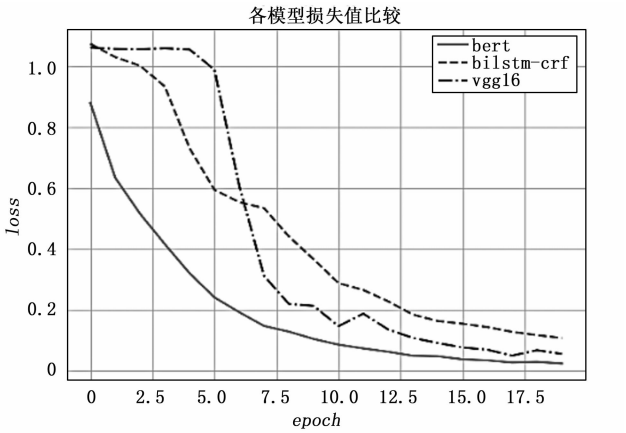

图 8　3 个模型训练过程损失值变化比较

基于上述实验比较，本文最终使用性能最优的 Bert 模型进行对剩余初始语料库进行知识抽取，得到多个实体信息和关系信息，完成知识抽取任务。

选择图数据库中的 Neo4J 软件来实现机床故障诊断知识的存储和可视化。图数据库的存储单元为三元组，三元组有“主体－关系－主体”“主体－属性－属性值”两种类型，为此，本文将前面得到数据转变成三元组类型，得到故障现象、故障操作、机床设备、操作与现象、现象与现象、现象与设备等三元组。

表 5　部分机床设备实体类文件内容格式表

id	name	label
1	电源	Part
2	驱动器	Part
3	显示屏	Partid

id 列为实体独特编号，用于区分不同实体，name 列为机床设备名，label 列为实体标签，机床设备实体标签为 Part，实体三元组存储的方式都是相同的，但与关系三元组存储方式存在差异。机床设备与现象的关系文件内容格式如表 6 所示。

表 6　机床设备与现象的关系文件内容格式表

from_name	relation	to_name
出现电源无法接通的故障	故障部位	电源
系统显示 ALM401	故障部位	系统
电源单元的熔断器 F14 已经熔断	故障部位	熔断器

其中，第一列记录故障现象实体类，第二列记录故障现象实体类与机床设备实体类之间的关系，第三列记录机床设备实体类。对应＜实体，关系，实体＞形式的三元组，第一个实体是图单向边的起始结点也就是表 4 的第一列，第二个实体对应第三列。

机床故障诊断的知识图谱在构建完成之后，实际运用的内部流程如图 10 所示。

将输入的故障信息，进行分句处理，使用训练好的知识抽取模型预测每句信息的类型，向 Neo4J 图数据库请求相应

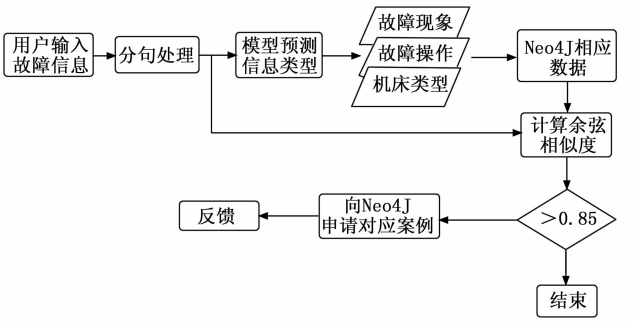

图 9　知识图谱工作流程图

```
similar_phenomenon:[‘坐标轴快速运动’,‘电压模块报警’,‘系统无法工作’,‘ALM413报警’]
reason_item:(‘连接不良引起跟随误差报警的故障’, 0.5)
```

图 10　高相似度信息

类型的数据列表，然后计算该句信息与数据列表每个数据的余弦相似度，将超过阈值 0.85 的数据保留，其余删除，之后根据保留的数据，再次向 Neo4J 图数据库请求与保留的数据周围相关的所有信息，最后整理给出反馈。关于与数据周围相关的所有信息，知识图谱以图的形式记录数据，通过单向边存储实体关系信息，结点存储实体信息，因此会将数据周围相连的结点信息以及单向边数据一并返回。

例如，用户输入“开机时，坐标轴快速运动，电压模块报警，系统无法工作”，系统通过计算会得到如图 11 相似现象，同时将故障原因一并返回。

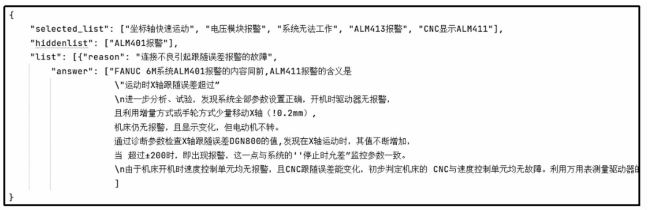

图 11　系统根据用户输入故障信息反馈结果

用户最终收到来自系统的反馈结果如图 12 所示，其中记录用户输入数据、潜在相关现象、故障原因以及合理的故障解决措施。

搭建好的机床故障知识图谱由于实体过多，导致整体显示时会较为模糊，因此可以通过部分设备以及相关属性来直观显示知识图谱。

在知识图谱中，可以从结点、单向边来看出实体之间的关系，也可以从结点的颜色来辨别故障实体。当机床实际出现故障时，可以通过语句查询故障现象、故障设备以及故障出现前的操作来快速定位具体的故障实体结点，知识图谱会进一步提供几种具体的故障原因和可行的维修方式。

4　结束语

本研究通过以机床为对象，从海量的数据当中提取核心信息，使用当前几种流行的命名实体识别模型进行比较，按照数据获取、数据处理、知识加工、知识抽取、知识存储与可视化表达这几个方面完成装配信息与故障诊断信息

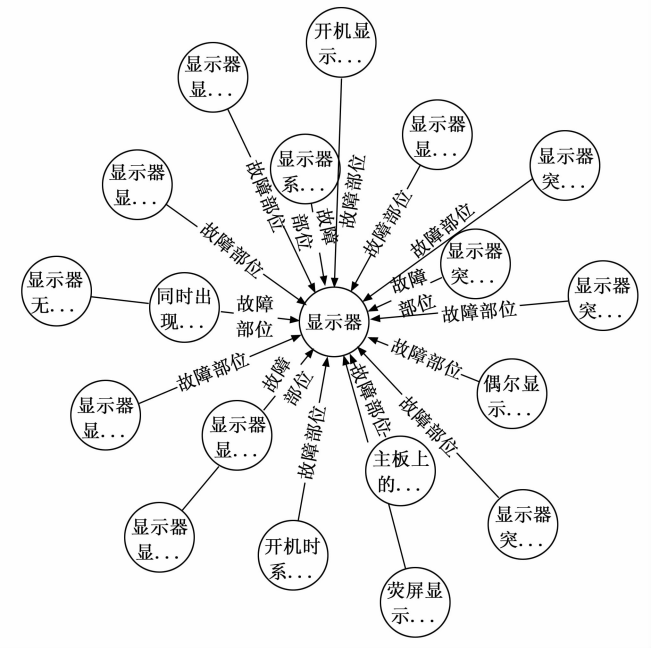

图 12 显示器的实体关系结点展示图

的知识图谱构建。核心研究内容如下：

1）以往关于机床的故障诊断信息大多分布在各个工厂的案例表当中，一个工厂出现的故障很可能就已经在其他工厂发生过。但由于不存在公开的标准故障诊断库，导致对机床的故障诊断效率不高。本文采用爬虫技术从互联网中的专业社区挖取故障诊断信息，另外解析专业书籍中的案例信息，分别提取机床型号、机床设备、机床故障现象、故障原因以及解决措施等信息。实现从大量口语化信息当中抽取核心信息，进一步使用模型训练，完成实体和关系信息的提取，最后实现故障诊断信息的存储。

2）在机床领域中，传统的命名实体抽取模型大多使用基于规则的方法，需要大量人力且效率不高，准确率依赖于处理人员的专业知识，且模型大多是表达单一连接关系，为线性模型。本文运用了多个当前高性能的命名实体识别模型，有 BiLSTM-crf 模型、Vgg16 模型、Bert 模型，分别对这几种模型进行测试分析，实验证明这几种模型性能均能实现功能，其中 Bert 模型具备最好的识别效果。

基于模型的机床故障案例命名实体抽取方法虽然可行且可以提高故障诊断效率，但是仍存在一些不足，例如：本文构建的机床故障诊断知识图谱可以对用户输入的设备故障现象给予反馈，但是故障诊断本身就充满未知性，许多隐含的故障难以解决，且随着案例不断增加，知识图谱需要经常维护以及更新，实体之间的歧义问题也会变得更加复杂，且缺乏知识推理环节，知识库还有很多潜在关系信息等待挖掘。

参考文献：

[1] 杜毓瑾. 数控机床 CNC 系统的故障诊断和维修研究 [J]. 机械管理开发，2012 (2)：29－30.
[2] 雷亚国，贾 峰，孔德同，等. 大数据下机械智能故障诊断的机遇与挑战 [J]. 机械工程学报，2018，54 (5)：94－104.
[3] 盛 博，邓 超，熊 尧，等. 基于图论的数控机床故障诊断方法 [J]. 计算机集成制造系统，2015，21 (6)：1559－1570.
[4] 林伟强. 基于循环神经网络的数控机床故障诊断研究 [J]. 机床与液压，2022，50 (5)：191－196.
[5] 马振林，于英杰. 基于 RBR 和 CBR 的故障诊断专家系统研究 [J]. 微计算机信息，2010，26 (4)：111－112.
[6] 朱传敏，周润青，陈 明，等. 故障树与案例推理在数控机床故障诊断专家系统中的应用研究 [J]. 制造业自动化，2011，33 (10)：21－24.
[7] 徐增林，盛泳潘，贺丽荣，等. 知识图谱技术综述 [J]. 电子科技大学学报，2016，45 (4)：589－606.
[8] BIZER C，LEHMANN J，KOBILAROV G，et al. DBpedia-A crystallization point for the web of data [J]. Web Semantics Science Services & Agents on the World Wide Web，2009，7 (3)：154－165.
[9] SUCHANEK F M，KASNECI G，WEIKUM G. YAGO：a large ontology from Wikipedia and WordNet [J]. Journal of Web Semantics，2008，6 (3)：203－217.
[10] 刘 浏，王东波. 命名实体识别研究综述 [J]. 情报学报，2018，37 (3)：329－340.
[11] RAU LF. Extracting company names from text [C] // Artificial Intelligence Applications，1991. IEEE，1991.
[12] ZHANG Z，HAN X，LIU Z，et al. ERNIE：enhanced language representation with informative entities [C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics，2019：1441－1451.
[13] 郭剑毅，李 真，余正涛，等. 领域本体概念实例、属性和属性值的抽取及关系预测 [J]. 南京大学学报：自然科学版，2012，48 (4)：383－389.
[14] 袁凯琦，邓 扬，陈道源，等. 医学知识图谱构建技术与研究进展 [J]. 计算机应用研究，2018，35 (7)：1929－1936.
[15] 尹昱东，王保建. 半结构装配数据的知识抽取及机床装配知识图谱构建方法研究 [J]. 制造技术与机床，2022 (11)：97－101.
[16] 刘 峤，李 杨，段 宏，等. 知识图谱构建技术综述 [J]. 计算机研究与发展，2016，53 (3)：582－600.
[17] 王 欢，朱文球，吴岳忠，等. 基于数控机床设备故障领域的命名实体识别 [J]. 工程科学学报，2020，42 (4)：476－482.
[18] YANG J，TENG Z，ZHANG M，et al. Combining discrete and neural features for sequence labeling [C] // Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics. Cham：Springer，2016：140－154.
[19] LIVADAS C，WALSH R，LAPSLEY D，et al. Using machine learning techniques to identity botnet traffic [C] // Local Computer Networks，Proceedings 2006 31st IEEE Conference on IEEE，2006：967－974.
[20] VENKATESAN K. Mining user profile exploitation cluster from computer program logs [J]. International Journal Innovative Research in Computer and Communication Engineering. 2015，3：1556－1561.