

基于知识图谱的电力设备故障诊断方法

王宁[†]

(国网浙江省电力有限公司湖州供电公司, 浙江 湖州 313000)

摘 要:现有的中文实体识别方法大多准确率较低,难以从中文技术文献中提取知识来构建高质量的知识图谱。因此提出了基于知识图谱的电力设备故障诊断方法,通过 BERT-BiLSTM-CRF 模型,建立电力设备故障的知识图谱。首先通过 BERT 模型对动态词向量进行预训练,然后引入 BiLSTM-CRF 模型,综合考虑文本的全局和局部特征,考虑数据标签的隐藏序列规则,提取实体之间的语义关系。最后,提取的知识以三元组的形式存储在 Neo4j 数据库中,并以图的形式进行可视化。以电力设备故障诊断的中文技术文献为实验对象,实验结果表明,该模型比传统方法更准确地识别和提取中文实体,从而更容易构建全面、准确的电力设备故障中文知识图谱。

关键词:知识图谱;故障诊断;BERT-BiLSTM-CRF;实体识别

中图分类号:TP273

文献标识码:A

Power Equipment Fault Diagnosis Method Based on Knowledge Graph

WANG Ning[†]

(Huzhou Power Supply Company, State Grid Zhejiang Electric Power Co., Ltd., Huzhou, Zhejiang 313000, China)

Abstract: Most of the existing Chinese entity recognition methods have low accuracy, and it is difficult to extract knowledge from Chinese technical literature to construct high-quality knowledge graph. Therefore, this paper proposes a fault diagnosis method of power equipment based on knowledge graph, and establishes the fault knowledge graph of power equipment through BERT-BiLSTM-CRF model. Firstly, BERT model is used to pre-train dynamic word vectors, and then BiLSTM-CRF model is introduced to comprehensively consider the global and local features of text and the hidden sequence rules of data labels to extract the semantic relationships between entities. Finally, the extracted knowledge is stored in the Neo4j database in the form of triples and visualized in the form of graphs. The experimental results show that the model can identify and extract Chinese entities more accurately than the traditional methods, so that it is easier to construct a comprehensive and accurate Chinese knowledge graph of power equipment fault diagnosis.

Key words: knowledge graph; fault diagnosis; BERT-BiLSTM-CRF; entity recognition

随着电网的不断发展,电力设备的种类和功能也越来越复杂。因此,设备的日常操作,如故障诊断和维护,高度依赖于工作人员的专业电力知识。然而,由于缺乏有效的知识提取、推理和应用,负责

运维的人员诊断电力设备故障主要依靠自己的经验^[1-2]。这种主观的方式不仅效率低下,而且难以保证准确性。事实上,电力行业经过几十年的发展,已经拥有了大量的中文技术文献,其中蕴含着

丰富的电气设备故障知识。如果能够将这些知识准确地提取出来,并以可理解的形式呈现给工作人员,那么故障诊断无疑会更加快速和准确^[8]。针对这一问题,一些学术和应用研究试图对这些知识进行提取、组织和展示,为电力设备故障的智能诊断提供更好的支持。其中,最引人注目的是知识图谱技术。

知识图谱的出现为电力行业提供了一种更加直观易用的数据存储方式,使碎片化的数据信息相互连接起来^[4]。因此,研究基于知识图谱的电力设备故障缺陷的实体关系提取与辅助处置具有重要的现实意义。文献[5]研究了电力设备缺陷文本知识图谱的构建方法,并基于构建的知识地图进行了电力设备缺陷文本检索与纠错的应用实验,取得了较好的效果。文献[6]对现有的输电线路螺栓缺陷分类方法进行了研究,总结了其局限性,利用螺栓与螺母之间的相关性提取其特征,建立相应的知识图谱,实现对螺栓缺陷的分类。文献[7]构建了适合于电力设备健康状态评估的电力设备健康管理知识图谱,并分析了该知识图谱的应用和需求。文献[8]以电力设备的数据信息为研究依据,提出了电力设备知识图谱的构建方法,达到了提高电力设备管理效率的目的。文献[9]在研究电力设备提供能源服务的工作模式后,提出了能源服务知识图谱的构建方法。

本文提出了一种基于 BERT-BiLSTM-CRF 模型的电力设备故障中文知识图谱创建方法。其新颖之处在于这种微妙的融合提高了中文实体识别和关系提取的准确性。首先,语言预训练模块 BERT 使用动态词向量进行初步实体识别。它不仅减少了下游任务的工作量,而且提高了中文实体识别的准确率。这是因为与静态词向量相比,动态词向量在中文实体识别中更有优势。例如,动态词向量可以在不同的上下文中表达不同的语义。此外,CRF 模块对标注序列进行约束和反转,解决部分标注偏差,计算整个标注序列的联合概率,充分保证了中文实体识别的准确性^[10]。

1 知识图谱构建

采用自底向顶的方式构建电力设备故障知识图谱,如图 1 所示。首先,收集了有关电力设备故障的各种文献,由于大多数文献是非结构化的,称之为图 1 中的“非结构化故障文献”,然后对它们进行预处理,并通过分词、词向量表示和关键字提取

构建本体词典。在此基础上,利用 BERT-BiLSTM-CRF 模型从预处理文献中识别和提取实体、关系等知识概念,然后使用依赖解析定义实体之间的关系模式,最后,用 Cypher 语言表达语义知识,并以三元组的形式存储在 Neo4j 数据库中。

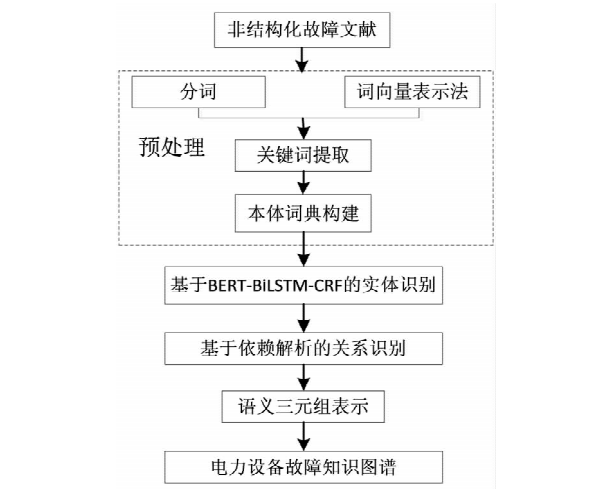


图 1 电力设备故障知识图谱的构建

2 基于 BERT-BiLSTM-CRF 模型的电力设备故障实体识别

2.1 文献预处理

对故障文献进行预处理,有助于提高后续实体识别的准确性,使更深层次的知识挖掘和知识图的构建更加全面。预处理包括分词、词向量表示和关键字提取,最后构建电力设备本体词典。

针对中文电力文献中大量同义词且歧义严重的问题,提出了隐马尔可夫模型-条件随机场(HMM-CRF)分词方法来提高中文分词的准确率。首先,利用基于通用词典的 HMM 模型对文献进行分词,对词进行排序,构建具有特征词频的高频词典;然后,使用基于 CRF 的分割模型对处理后的文献进行再次分割,导入高频词典;最后,获得了精度较高的分割结果,该技术还可以完成词性标注和词性停词的任务。

为了将文本数据转换为数字向量,使计算机能够识别,使用 Word2vec 模型来表示文本数据。Word2vec 模型包含连续词袋模型(CBOW),它将高维离散的一维向量转化为低维向量,如图 2 所示。该模型通过计算词向量之间的余弦相似度,从同义词中识别出相似度较高的词串,在语料库中获得的词向量也可以作为后续实体识别模型的输入。

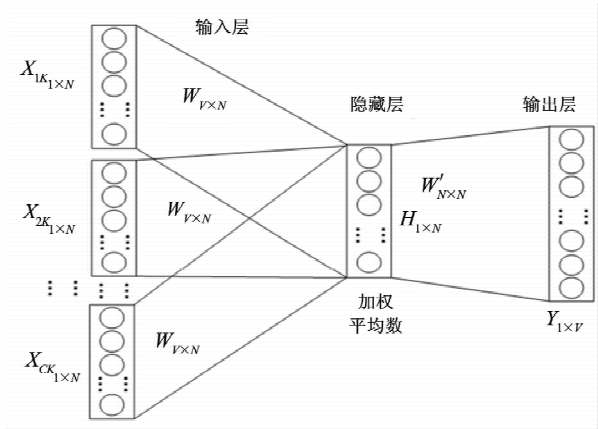


图 2 CBOW 模型示意图

在图 2 中,词向量 X_i 的维数 N 等于语料库的整个字典的大小, C 表示指定窗口的数量, W 和 W' 分别表示权矩阵, 隐层向量 $H_{1 \times N}$ 是输入层和矩阵的加权平均数, 使用 Softmax 分类器得到维数 v 的向量, 每个维数代表词库中的一个词, 概率最大的索引词是预测的中间词。

为了提高后续电力设备实体识别的准确性, 本文从故障文献中提取关键字, 构建电力设备本体字典。首先提取文本中的词频和逆文档词频, 计算频率权值; 然后, 通过计算平均信息熵来衡量整个文献中词汇分布的均衡程度; 最后, 基于频率权值均值和平均信息熵提取高频关键词。通过人工筛选去除不相关词, 构建本体词典, 有助于提高 BERT-BiLSTM-CRF 模型标注阶段的准确率。

2.2 基于 BERT-BiLSTM-CRF 模型的实体识别

传统的深度学习实体识别方法, 如基于 LSTM 模型的方法, 没有考虑上下文信息特征。BiLSTM 模型考虑了上下文信息, 但忽略了标注序列的局部特征对模型效率的影响, 导致实体识别的准确率较低。因此提出了 BERT-BiLSTM-CRF 模型来准确提取电力设备的实体。首先通过 BERT 模型对动态词向量进行预训练, 全面学习句子的句法和语义特征, 同时, BiLSTM-CRF 模型综合考虑文本的全局和局部特征, 考虑数据标签的隐藏序列规则, 预测并输出最佳序列。模型包含三层, 如图 3 所示。

模型的第一层是 BERT 层, BERT 层通过多层神经网络进行特征提取和训练, 并将输入文本转换成词向量, 使 BiLSTM 层能够学习上下文特征。BERT 模型将输入序列转换为标记、分段和位置三个特征的综合嵌入, 然后输入到模型中, 如图 4 所示。

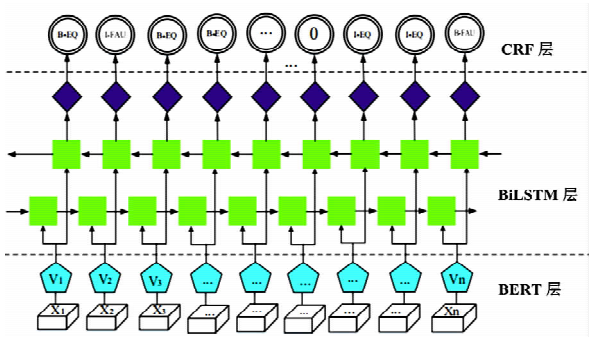


图 3 BERT-BiLSTM-CRF 模型示意图

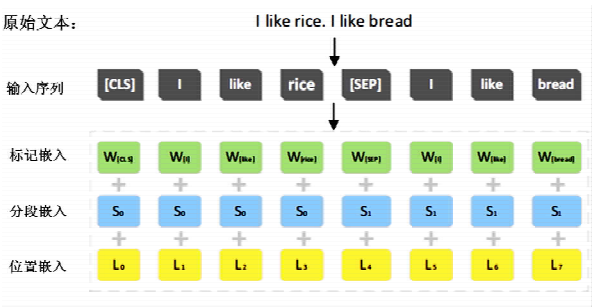


图 4 BERT 模型的输入表示

BERT 模型中最重要的模块是用于特征提取的双向 Transformer 编码结构, 它使用自关注机制和全连接层对输入文本进行建模, BERT 模型训练的动态词向量可以在不同的语境下表达不同的语义。与传统语言训练模型得到的静态词向量相比, 得到的词向量不仅包含词本身的含义, 还包含词的上下文, 还可以捕捉句子层面的隐含特征。第二层是自动提取句子上下文特征的 BiLSTM 模型, 每个 BiLSTM 单元的输入是一个动态词向量序列, 然后, BiLSTM 单元学习如何提取句子的局部特征。最后, 前向 LSTM 模型输出隐藏状态序列, 后向 LSTM 模型根据句子序列拼接所有隐藏状态序列, 得到完整的隐藏状态序列。相关数据可由式 (1)~式 (5) 得到:

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

$$f_t = \delta(W_f \times [h_{t-1}, x_t] + b_f) \tag{2}$$

$$O_t = \delta(W_o \times [h_{t-1}, x_t] + b_o) \tag{3}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tan(W_c \times [h_{t-1}, x_t] + b_c) \tag{4}$$

$$h_t = O_t \times \tanh(C_t) \tag{5}$$

式 (1)~式 (5) 中, i_t 、 f_t 、 O_t 分别表示每个 LSTM 单元的三个门控单元: 输入门、遗忘门、输出门。 C_t 表示 t 时刻输出层的输出状态, h_t 表示 t 时刻隐藏层的输出状态。 x_t 表示 t 时刻的输入。 δ

() 为激活函数, $\tanh(\cdot)$ 为双曲正切激活函数。 W_i 、 W_f 、 W_o 表示隐藏状态向量 h_t 与输入向量 x_t 的权重矩阵, b_i 、 b_f 、 b_o 、 b_c 表示偏移向量。

将 BiLSTM 模型的输出序列导入分类器, 分类器选择最大概率值作为该序列的标签输出。由于 Softmax 分类器输出的独立性, 如果不考虑句子的局部特征, 输出序列将是无序的。如图 5 所示, 可以看出 B-EQ 和 B-FAU 都错在 I-EQ 或 I-FAU 之后。这导致了训练模型的准确率较低。因此, 提出了综合考虑句子隐含序列规则的 CRF 模型。

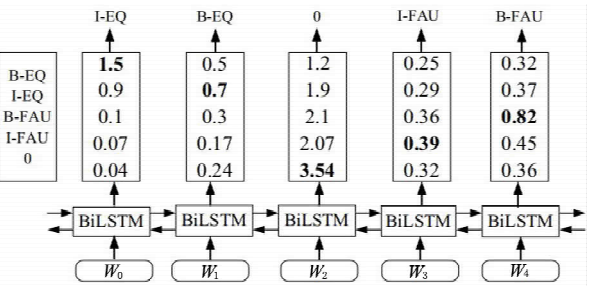


图 5 识别序列的 BiLSTM 序列图

最后一层是 CRF, 它是一个用无向图表示的联合概率分布的图模型。将局部特征归一化为全局特征, 通过计算整个序列的概率分布来解决部分标记偏差问题。在模型中加入 CRF 层后, 实验结果证明可以得到全局最优解。同时, CRF 模型在训练数据时可以获得标签的隐藏约束规则。有了这些约束规则, 实体识别的准确性将大大提高, 如图 6 所示。

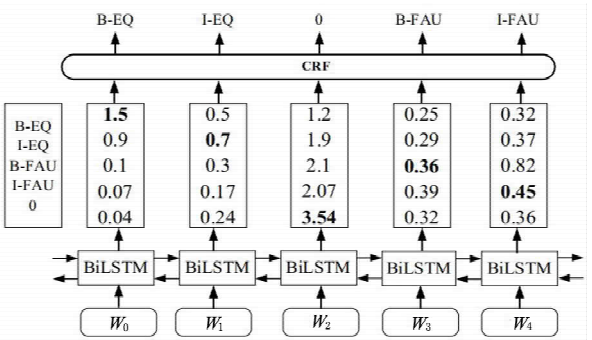


图 6 BiLSTM-CRF 序列图

2.3 关系提取

实体之间的关系是知识图谱中的关键环节, 它直接影响到知识图谱的知识多样性, 电力故障文献关系提取的目的是识别和获得被提取实体之间的语义关系, 然而, 电力行业并没有固定的结构化关系规范, 因此基于依赖解析确定被提取实体之间的关系。

依存句法分析是通过对句法关系的识别和定位, 来分析句子成分之间的依存关系。首先, 通过语义角色识别提取主题和核心谓词; 然后, 通过依存句法分析, 找到与核心谓词意义相关的宾语和主语; 最后, 通过依赖句法分析得到电力故障文献中的相关依赖关系, 并基于本体结构得到实体关系。

2.4 知识存储与可视化

将识别出的实体和关系进行知识聚合形成语义三元组, 用密码语言表示, 并将知识导入图形数据库 Neo4j 进行知识存储和知识边缘可视化。Neo4j 是一个稳定成熟的图形数据库, 具有高性能、可用性和可扩展性。通过遍历 Neo4j 数据库, 可以高效快速地检索知识数据。

3 实验分析

3.1 实验数据

为了验证从电力技术文献语料库中构建电力设备故障知识图谱的可行性, 提出基于 BERT-BiLSTM-CRF 的模型用于实体识别的准确性, 进行了四次对比实验。本文选取 2003—2022 年 CNKI 收集的 1200 篇关于电力设备故障的文档, 随机抽取 450 篇文档, 按照 8:2 的比例分为训练集和测试集, 使用 BIO 对数据集进行标记, 并通过 YEDDA 工具完成标记, 电力技术文献标注见图 7。

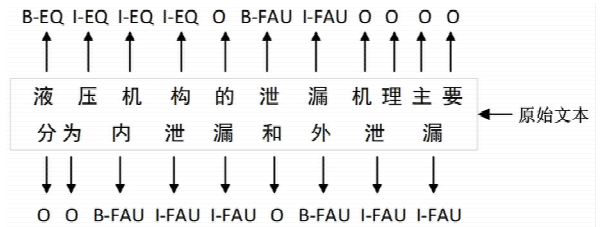


图 7 电力技术文献 BIO 标签

图 7 中的实框是原始文本, 我们在每个字符上都放了一个标签。其中“FAU”为故障实体, “EQ”为设备实体, “ORG”为组织实体; “B”表示实体开头的字符, “I”表示组成实体的非开头字符, “O”表示非实体字符。

3.2 实验参数

为了使 BERT-BiLSTM-CRF 模型在实体识别任务中达到更高的精度, 需要确定更好的模型参数。因此对不同的 epoch 值和向量维度进行了对比实验。当 epoch 值分别为 40、60、80 时, 进行不同词向量维数下的对比实验, 结果如图 8 所示。

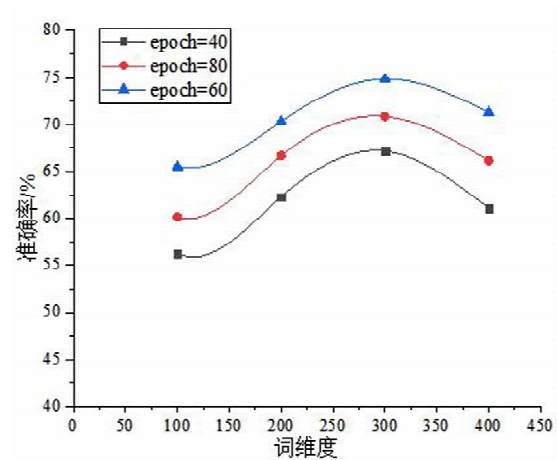


图 8 不同词维度的准确性比较

从图 8 的对比实验结果可以看出,当 epoch 为 60 时,模型的整体效果最好。同时,BERT-BiLSTM-CRF 模型的效果在一定范围内随向量维数的增加而增加,而不是一直增加。当维度为 300 时,识别精度最高;当维度为 400 时,精度低于 300,即矢量维数与模型效果之间不存在线性关系。因此,需要根据具体任务要求,通过对比实验确定模型参数。模型最终参数如表 1 所示。

表 1 模型参数

参数	值
词维度	300
梯度剪切	5.0
Dropout	0.5
批量大小	1000
epoch	60
隐藏维度	100
单元数	256
学习率	0.001

3.3 准确性验证

本文进行了四个实验,比较和验证了模型在电力设备故障文本中识别中文实体的效果。实验 1 为 LSTM 模型,实验 2 为 BiLSTM 模型,实验 3 为 BiLSTM-CRF 模型,实验 4 为本文 BERT-BiLSTM-CRF 模型。中文实体包括三类:设备、故障和组织,采用五重交叉验证,评价指标为正确率、召回率、 F_1 。四次实验结果如表 2 所示。

表 2 实体识别比较

%

模型	指标	设备实体	故障实体	组织实体
BERT-BiLSTM-CRF 实验 4	Accuracy	81.32	79.17	88.23
	Recall	77.08	75.25	91.67
	F_1	78.32	77.16	89.80
BiLSTM-CRF 实验 3	Accuracy	74.09	72.03	86.07
	Recall	75.88	78.16	91.03
	F_1	72.95	70.02	91.84
BiLSTM 实验 2	Accuracy	65.78	63.94	88.14
	Recall	63.10	60.83	86.15
	F_1	64.42	61.39	82.74
LSTM 实验 1	Accuracy	59.18	57.64	86.17
	Recall	54.29	56.19	85.54
	F_1	56.79	55.31	84.04

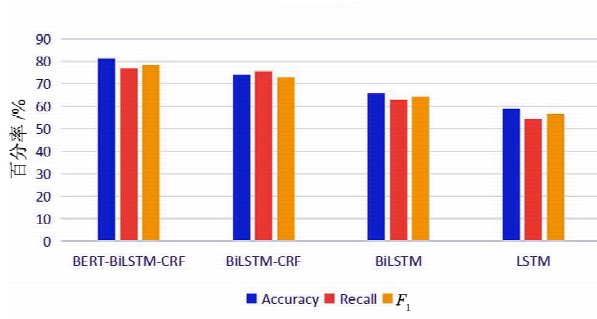
对比四种模型对三种类型实体的识别效果,BERT-BiLSTM-CRF 模型的识别效果高于其他三种模型,且差异最显著的是设备实体识别的准确率。这说明经过 BERT 层处理后可以得到语境化的动态词向量,在中文实体识别方面比静态词向量更有优势,在处理长距离依赖信息句方面效果很好。此外,组织实体识别和故障实体识别的准确率也高于其他三种模型,因此,通过与其他三种模型的比较,验证了 BERT-BiLSTM-CRF 模型在识别设备实体方面的准确性。

如图 9 所示,可以更形象地看出四种模型的实验指标对比效果。

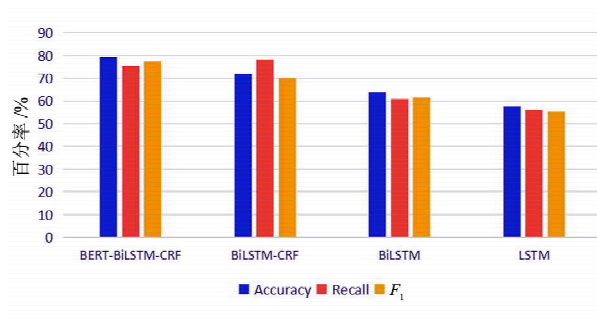
3.4 训练效率对比

模型训练是参数更新的过程,参数更新是通过迭代实现的,图 10 显示了迭代次数与四个模型的 F_1 之间的关系。与 LSTM 模型相比,BiLSTM 模型包含正向和反向双向信息。BiLSTM 可以更好地捕获双向语义依赖,因此 BiLSTM 模型的效果比 LSTM 模型有显著提高。与 BiLSTM 模型相比,BiLSTM-CRF 在识别效率上保持了持续稳定

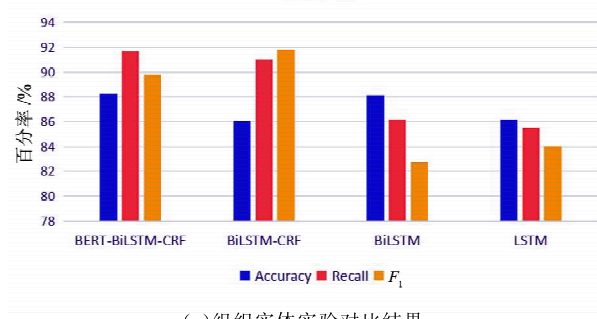
的提高,最终使模型收敛速度更快。这反映了 CRF 层显著提高了模型的收敛性。CRF 层计算 BiLSTM 输出的每个字符标签序列的损失,以保证模型更好的收敛性。与其他三种模型相比,BERT-BiLSTM-CRF 模型在初始训练阶段可以更快地达到更高的水平,这表明 BERT 模型在迭代训练过程中可以快速准确地识别上下文语义特征。BERT-BiLSTM-CRF 模型可以获取短语级特征信息和学习句法结构特征信息,从而捕获整个句子的语义信息。



(a)设备实体实验对比结果



(b)故障实体实验对比结果



(c)组织实体实验对比结果

图 9 不同模型实体识别对比结果

如图 10 所示,BERT-BiLSTM-CRF 模型可以在更短的时间内达到比其他三个模型更高的 F_1 分数。在模型迭代到一定次数并达到收敛后,BERT-BiLSTM-CRF 模型在处理本文的任务时获得了最高的 F_1 分数。这表明 BERT-BiLSTM-CRF 模型具有较高的训练效率,可以快速地用于

电力设备故障的中文实体识别,并且效果也是最好的。

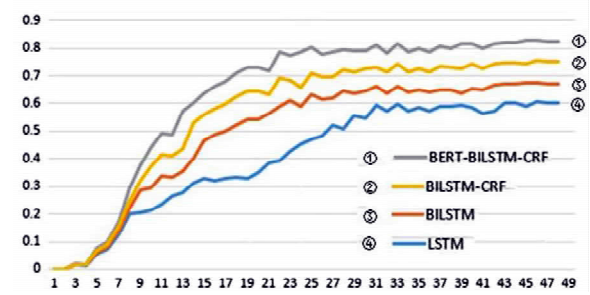


图 10 四种模型在设备实体识别训练效率上的比较

3.5 电力设备故障中文知识图谱的构建

基于上述实体识别方法,可以对电力缺陷域中的实体进行识别,通过对电力文献中四种依赖句法结构的识别,可以提取出六种依赖关系,然后对提取的关系进行过滤和清洗,去除重复部分,最后得到实体之间的关系,表 3 显示了语义关系的一些示例。

表 3 实体语义关系的例子

实体 A	实体 B	关系
主变压器	3 号变压器	包含
变压器	平台区域	范围
馈线机	变压器	连接
电压互感器	局部放电	发生
损坏	严重	程度
局部放电	跳闸	原因

通过实体识别和实体关系提取,将电力设备故障的语义三元组用密码语言表示,导入 Neo4j 数据库。Neo4j 数据库的知识存储和知识可视化是通过密码语言实现的,完成了电力设备故障知识图谱的构建,部分知识图谱如图 11 所示。

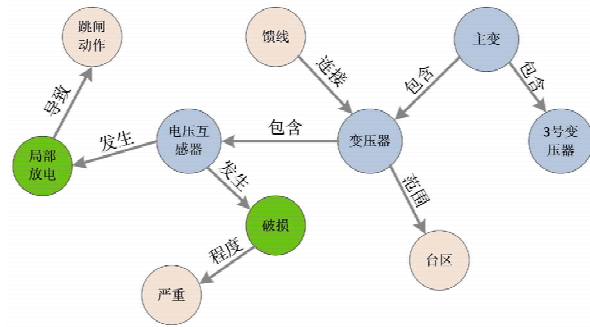


图 11 部分电力设备故障知识图谱

4 结 论

为了提高电力设备中文技术文档故障实体识别的准确性,提出了 BERT-BiLSTM-CRF 模型。该模型的新颖之处在于通过三层的微妙融合,从文献中学习上下文信息,并计算整个标注序列的联合概率,这样可以完全获得训练数据的相邻信息,提高中国电力设备实体识别的准确性。实验结果表明,与基于 BiLSTM-CRF 模型、BiLSTM 模型和 LSTM 的传统方法相比,基于 BERT-BiLSTM-CRF 模型的方法提高了准确率和训练效率。因此,本文的模型和方法对于构建全面、准确的电力设备故障中文知识图谱具有重要的指导意义。

参考文献

[1] 朱愈,张耘.融合 CANOPY 的 K-MEANS 算法在电力设备故障检测中的应用研究[J]. 自动化与仪器仪表,2024(4):62—66.

[2] 彭曙蓉,刘登港,何洁妮,等.基于改进 YOLOV4 的架空线路电力设备故障检测[J]. 电力科学与技术学报,2023,38(5):169—176.

[3] 刘炎昌.基于知识图谱的电力通信设备故障诊断方法研究及

应用[D]. 吉林:东北电力大学,2023.

[4] 陈滇斐,章黄勇,马宏忠,等.基于深度学习的电力设备故障诊断方法研究综述[J]. 电气自动化,2022,44(1):1—2+6.

[5] GENG J, CAO B, YE H, et al. Web service recommendation based on knowledge graph convolutional network and doc2vec[C]//2020 IEEE World Congress on Services (SERVICES). IEEE, 2020: 95—100.

[6] PALUMBO E, MONTI D, RIZZO G, et al. Entity2rec: property-specific knowledge graph embeddings for item recommendation[J]. Expert Systems with Applications, 2020, 151: 113235.

[7] UNFENG Q, LIN P, AIFUA Z. An Improved electric equipment fault prediction based on recursive multi-parameter prediction algorithm[J]. Journal of Intelligent & Fuzzy Systems, 2024, 46(4):8025—8035.

[8] PLOTNEK J J, SLAY J. Power systems resilience: Definition and taxonomy with a view towards metrics[J]. International Journal of Critical Infrastructure Protection, 2021, 33: 100411.

[9] LISTL F G, FISCHER J, BEYER D, et al. Knowledge representation in modeling and simulation; a survey for the production and logistic domain[C]//2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, 2020, 1: 1051—1056.

[10] 周保先.面向电力通信设备故障信息的知识图谱构建方法研究[D]. 天津:天津科技大学,2022.