文章编号：2096-5125 (2020) 06-0607-011 中图分类号：TP391.1；TM41 文献标志码：A
DOI：10.19705/j.cnki.issn2096-5125.2020.06.008

面向变压器智能运检的知识图谱构建和智能问答技术研究

张敏杰[1]，徐宁[2]，胡俊华[3]，王宇飞[1*]，李晨[2]，徐剑波[4]，张诗玉[4]
（1. 善智互联（北京）网络科技有限公司，北京市 海淀区 100089；
2. 国网浙江省电力有限公司电力科学研究院，浙江省 杭州市 310014；
3. 国网浙江省电力有限公司，浙江省 杭州市 310007；4. 北京大数据研究院，北京市 海淀区 100089）

Knowledge Graph Construction and Intelligent Question Answering for Transformer Operation and Maintenance

ZHANG Minjie[1], XU Ning[2], HU Junhua[3], WANG Yufei[1*], Li Chen[2], XU Jianbo[4], ZHANG Shiyu[4]
(1. Kindi (Beijing) Network Technology Co., Ltd., Haidian District, Beijing 100089, China;
2. Electric Power Research Institute of State Grid Zhejiang Electric Power Co., Ltd., Hangzhou 310014, Zhejiang Province, China;
3. State Grid Zhejiang Electric Power Co., Ltd., Hangzhou 310007, Zhejiang Province, China;
4. Beijing Institute of Big Data Research, Haidian District, Beijing 100089, China)

Abstract: During transformer-equipment operation and maintenance, power companies encounter problems such as generation of difficult-to-use unstructured text data, full-caliber data that are difficult to integrate deeply, and application of shallow equipment knowledge application. Using artificial-intelligence techniques, such as the semantic web, knowledge map, and natural language processing, this study investigates the key technologies used for intelligent servicing of equipment. Accordingly, an intelligent technology framework for transformer operation and maintenance is proposed. This framework comprises three components—intelligent unstructured-text recognition and extraction, device-centered knowledge representation and storage, and intelligent device knowledge application. Three deep-neural-network-based intelligent models for device semantic extraction, device semantic-similarity calculation, and intelligent question answering are proposed to summarize three scenarios—review of the transformer running-status evaluation report, flexible equipment information querying, and auxiliary diagnosis. Finally, directions for future research on knowledge-based equipment intelligence technologies are identified.

Keywords: intelligent operation and maintenance; knowledge graph; transformer fault diagnosis; intelligent question answering

摘 要：针对电力公司在开展变压器设备运检过程中存在的非结构化文本数据难以利用、全口径数据难以深度融合、设备知识应用深度较浅等难题，基于语义网、知识图谱、自然语言处理等人工智能技术，对开展设备智能管理的关键技术进行了研究，提出了支撑变压器智能管理的智能技术框架，包括非结构化文本智能识别与提取、以设备为中心的设备知识表示与存储、设备知识服务应用三部分；形成了设备语义提取模型、设备语义相似度计算模型、基于深度神经网络的智能问答模型等三个智能模型；总结了该技术在变压器设备状态评价报告自动化审核、设备信息灵活查询、基于设备故障知识的辅助诊断三个场景的应用成果；提出了基于知识的设备智能技术下一步研究的方向。

关键词：智能运检；知识图谱；变压器故障辅助诊断；智能问答

0 引言

在电力设备运检领域开展人工智能研究的目标之一是实现机器理解设备运行检修专家的经验知识[1]，包括理解电力领域专有的语言、融合处理设备运检相关结构化和非结构化的全口径数据等[2]，从而让机器利用知识去丰富智能化运检手段。目前研究热点包括设备文本数据挖掘、故障推理与辅助诊断、设备缺陷检索等方面。

基金项目：国家电网公司科技项目（5500-202019090A-0-0-00）。
Science and Technology Foundation of SGCC (5500-202019090A-0-0-00).

设备文本数据挖掘是指从电力设备庞杂的文本数据中提取出电力领域知识，如电力设备缺陷知识。对电力设备领域的数据进行知识挖掘只是知识工程的第一步，最终目标是为了有效提升设备管理的效能[3]，如通过对缺陷文本的智能分类[4-5]实现变压器设备相关的检索、推理、故障诊断[6-8]。目前此类研究主要基于传统的专家系统思路，知识构建及更新成本较高，难以应对电力系统的大规模知识抽取、低成本应用等要求。近年来人工智能技术蓬勃发展，尤其是Google公司提出的知识图谱技术、基于变换器的双向编码器表征量模型（Bidirectional Encoder Representations from Transformers, BERT）等，为解决上述问题提供了新方向。

本文借鉴文献[9-11]提及的知识图谱相关的构建、查询等关键技术，结合作者在电力设备领域的知识图谱关键技术研究与实践成果，提出了一种支撑变压器智能管理的设备知识图谱技术框架，研究了基于知识开展变压器设备智能运检的关键技术，形成了设备语义提取模型等智能模型，总结了该技术在变压器故障报告自动化提取、设备信息灵活查询等场景的应用成果，并提出了基于知识的设备智能运检技术下一步研究的方向，为实现机器理解设备语义[12-13]提供了可行性。

1 变压器设备运检智能化面临的挑战

电力公司设备运检工作正逐步向智能化方向演进，目前主要面临三方面挑战：

1）电力公司在开展设备运行检修过程中积累了大量长文本数据，包括设备故障报告、试验检测报告、标准导则等，这些文本数据实现了文档归集，但历史数据中蕴含的知识价值未得到深入挖掘。

2）设备的全口径数据包括通过关系型数据库（relational database, RDB）保存的结构化数据和非结构化数据，两类数据之间关系尚未打通，难以形成以设备为中心的知识体系。

3）在设备运检中存在大量的知识，如设备检修标准、导则、设备故障案例等，对这些共性知识的应用形式以关键字检索为主，基于知识的智能决策、设备语义判断、知识驱动的故障诊断等知识应用尚处于探索阶段。

针对上述挑战，本文研究了变压器设备知识图谱构建技术及基于模型的智能问答技术[14]，提高电力设备运检智能化水平。

2 设备知识图谱关键技术研究

2.1 设备知识图谱构建技术

知识图谱是一个将实体和属性通过关系进行联结和组织的知识网络[7]，由“实体–关系–实体”或“实体–属性–属性值”三元组构成。

知识图谱分为开放域和封闭域2类[15]。本文研究的是面向电力变压器设备智能运检领域的知识图谱构建技术，属于封闭域。知识图谱构建过程如图1所示，由知识抽取、知识融合、知识存储3部分构成。

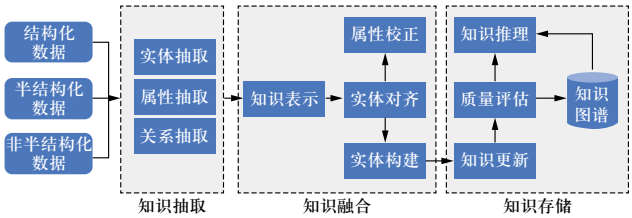


图 1 变压器设备知识图谱构建过程

Fig. 1 Knowledge graph construction process for the transformer

2.1.1 变压器结构化数据知识抽取技术

变压器设备相关的结构化数据包括设备台账数据、设备缺陷数据、工单数据等。这些数据通过RDB进行存储，具有明显的结构化特点，通过结构化数据源表结构与知识图谱本体概念的字段映射，快速将RDB数据转换为资源描述框架（resource description framework, RDF）数据。

2.1.2 变压器非结构化文本数据知识抽取技术

通过命名实体识别（named entity recognition, NER）从变压器设备长文本（如故障报告、试验数据等）中抽取实体和关系。本文综合采用基于规则模板的抽取+基于智能模型的方式实现NER。

1）基于规则模板的识别。适用于从具有较为固定格式或者结构的描述文本中抽取知识的场景，如“黄泥头110 kV变电站#2主变跳闸原因分析报告”等，变电站名称与变压器名称呈现出明显的模板性特征，针对此类文本语句，基于Python正则表达式（regular expressions, RE）的脚本模板抽取，效率、准确度高，成本低。

2）基于智能模型的识别。针对电力公司历年变压器故障报告数据，分别利用条件随机场（CRF）、双向长短期记忆网络（bi-directional long short-term memory, BiLSTM）、BERT-BiLSTM-CRF三种模型进行抽取[14]，综合对比三种模型评价指标准确率（precision, P）、召回率（recall, R）以及F_1值来优选采

用的模型。其中，BERT-BiLSTM-CRF模型结构如图2所示。

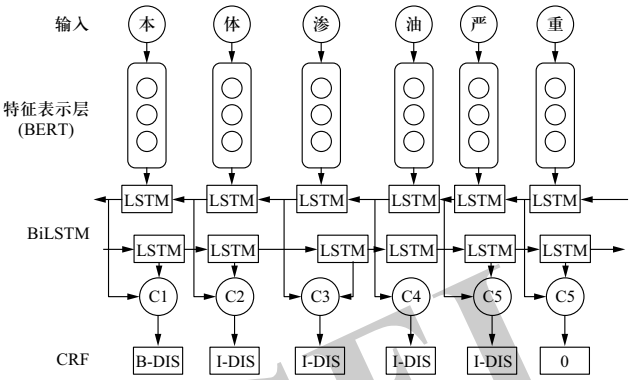


图 2　BERT-BiLSTM-CRF命名实体识别模型结构

Fig. 2　Architecture of BERT-BiLSTM-CRF model

在图2中可看出，BERT-BiLSTM-CRF命名实体识别模型通过使用BERT模型作为特征表示层加入到BiLSTM模型中，对每一个输入语句序列应用一个向前和向后的LSTM网络，并且这2个网络连接着共同的输出层，这种结构可给输出层提供输入语句序列完整的上下文信息[16]，最后经过CRF模型，有效地考虑了序列前后的标签信息。

2.1.3　变压器知识融合技术

对来自设备管理等系统的结构化数据和非结构化文本数据源的变压器知识完成抽取后，将知识进行存储时，需解决2类知识的冲突融合问题。

以变压器实体名称、数值属性、关系属性作为特征量，计算2个实体的语义相似度。在开展实体相似度计算前，需对实体的属性值进行数据预处理，对变压器相关实体的枚举型属性值进行归一化处理。如运检工作中变压器的电压等级值常见表述包括“110 kV”“110千伏”“22万伏”等；变压器故障现象包括“本体严重渗油”“本体漏油严重”“本体漏油成流状”等，对数据需归一化处理，调整为“本体严重渗油”。实体相似度计算通过对实体特征量的文本相似度计算加权得出。2个实体特征量相似度计算如式（1）所示。

$$S_{\mathrm{im}}(A,B)=\alpha S_{\mathrm{im}}\left(A_0,B_0\right)+\beta\sum_{i=1}^{n}\left(A_i,B_i\right)+\gamma\sum_{j=1}^{m}\left(A_j,B_j\right) \quad (1)$$

式中：A_0，B_0指的是A实体和B实体的实体名称；A_i，B_i指的是A实体与B实体的数值属性值；A_j，B_j指A实体和B实体的对象属性值；$S_{\mathrm{im}}(A,B)$指的是2个属性值的语义相似度；$\alpha+\beta+\gamma=1$，其中α、β、γ分别代表了实体名称相似度、实体数值属性值相似度、实体对象属性值相似度的权重。

实体属性值分为数值型、集合型、文本型3种。

对于数值型的属性，使用如式（2）的计算公式计算相似度。

$$D\left(d_i,d_j\right)=\frac{\left|d_j-d_i\right|}{\max(d_n)-\min(d_n)} \quad (2)$$

式中：D的值域范围为0到1，d_i与d_j之间差距越大，D值越大，表示两者之间的相似度越低。

对于集合型的属性，通过计算Jaccard相似度判断2个集合的相似度，如式（3）所示。

$$Jaccard\left(A,B\right)=\frac{\left|A\cap B\right|}{\left|A\cup B\right|} \quad (3)$$

式中：对于2个集合A、B，Jaccard相似度的值域范围是0到1，值越大，2个集合相似度越高。

对于文本型的属性，通过开源jieba分词工具进行分词，使用词向量计算的工具word2vec训练形成词向量模型，将实体的属性值转换为词向量，通过余弦相似度计算不同实体同一属性值的向量的相似度，如式（4）所示。

$$\cos S_{\mathrm{im}}=\frac{\sum_{i=1}^{n}\left(A_i\times B_i\right)}{\sqrt{\sum_{i=1}^{n}\left(A_i\right)^2}\times\sqrt{\sum_{i=1}^{n}\left(B_i\right)^2}} \quad (4)$$

式中：余弦相似度值域范围是0到1，值越大，则2个实体相似度越高。

2.1.4　知识存储技术

目前，知识图谱存储技术主要包括基于RDF模型的三元组数据库和基于属性图模型的图数据库[17]。在基于知识的设备智能应用场景中，结合考虑标准开放性、语义分析需求、支撑组件丰富度、易扩展性等要素，本文采用语义网RDF模型作为变压器设备知识图谱的知识表示模型，综合使用支持RDF存储的三元组数据库及开源MongoDB数据库组件实现对变压器知识图谱数据的存储。

2.2　变压器领域智能问答技术

变压器设备知识分为知识图谱（knowledge graph, KG）三元组、静态问答对（question/answering, Q/A）、网页文档等[18-19]。变压器领域智能问答的关键技术分为2项口语理解技术任务：意图识别和槽位提取。前者是对用户输入语句的分类任务，通过实现一个函数，输入是一个语句的n个词汇，$x=[x_1,\cdots,x_n]$，输出该语句对应的意图标签c；后者是序列标注任务，在这

个任务中，通过实现一个函数，输入是一个语句包含的n个专业词汇，$x=[x_1,\cdots,x_n]$，输出问句槽位的标注序列，$y=[y_1,\cdots y_n]$，其中，y_i是词汇x_i对应的槽位标签。

本文采用的关键技术思路参考了文献[20-22]，通过综合应用神经网络技术与正则表达式，解决用户口语化输入的语句与变压器KG中实体进行链接的问题[20]。从用户输入到最终返回答案的解析过程如图3所示。

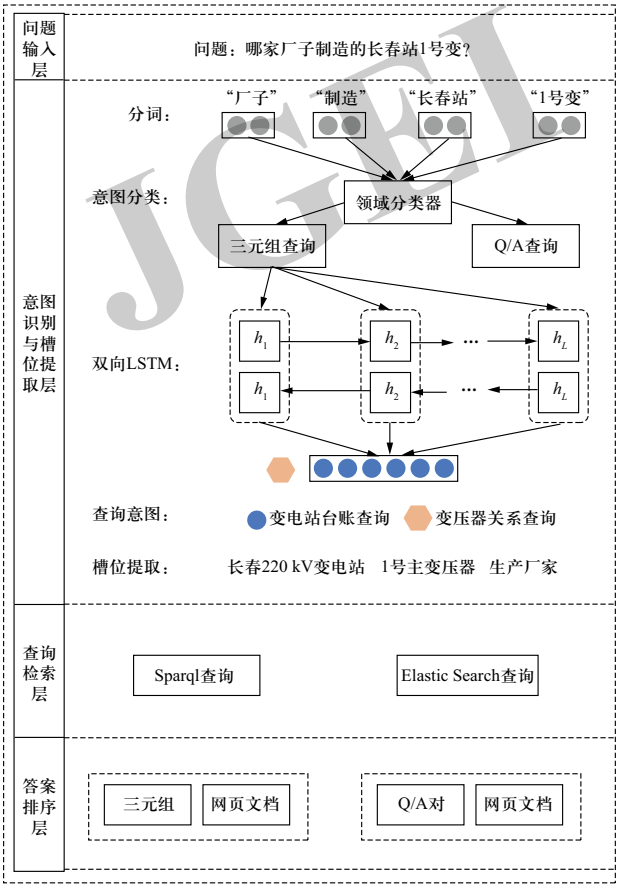


图 3 智能问答解析流程

Fig. 3 Intelligent question-answering analysis process

整个流程从上而下分为4层，上层的输出作为下层的输入，分别是问题输入、意图识别与槽位提取、查询检索、答案排序。

2.2.1 问题输入层

问题输入层用于接收来自系统界面的用户输入，这层的特点是用户对问题的描述呈现出模糊性、口语化、简写、同一问题存在各种近义表述、多次交互输入等问题。例如，用户要查询“长春站1号主变生产厂家是哪里”，该问题也可能输入为“长春220 kV变电站1号主变压器生产厂家是哪家”、“哪家厂子制造的长春站1号变”等近义问句。

2.2.2 意图识别与槽位提取层

问题输入层转发后的问题文本，根据意图类型分类器，判断用户的意图类型，通过BiLSTM模型实现意图识别与槽位提取，识别用户意图、提取问句槽位信息。

1）对输入问句运用正则表达式解析，提取特征量，过程中加入近义词的解析，如“哪家厂子制造的长春站1号变”，完成分词后，将“制造”的近义词“生产”，“厂子”的近义词“生产厂家”加入到问句模板的判断中。

2）问句意图识别过程如图4所示。

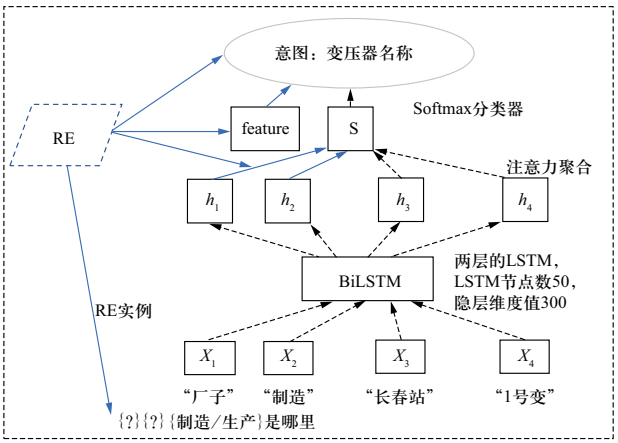


图 4 意图识别过程

Fig. 4 Intent recognition process

在图4中，对随机初始化的正则表达式标签嵌入进行平均，以构建聚合嵌入作为神经网络（neural network, NN）输入，最终作为Softmax分类器的输入。基于Softmax分类器，实现问句意图的识别，识别出意图为“变压器关系查询”。

3）槽位提取模型：对输入问句进行分词，将分词结果采用通用领域的预训练词向量进行嵌入，输入到BiLSTM槽位提取模型，BiLSTM输出序列，逐项进入Softmax分类器，完成词的槽位的类别识别，再对输出槽位类别序列进行拼接过滤，最终输出识别的槽位结果，如图5所示。

BiLSTM模型通过双向连接和注意力权重，有效解决上层文本输入的模糊性，根据目标向量输出较高置信度的概念、实体、属性。

2.2.3 查询检索层

查询检索层根据意图识别层传递的概念、实体、属性，构建查询子图，将查询子图转换为RDF开发的查询语言和数据获取协议查询语句（SPARQL protocol and RDF query language, SPARQL）或者搜索

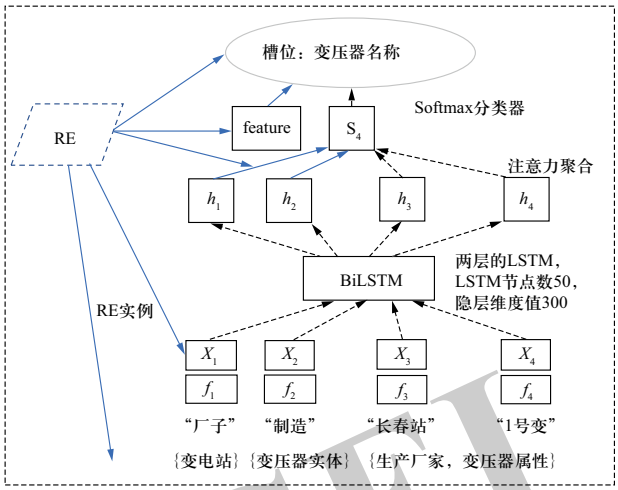


图 5　槽位填充（预测“变压器”槽位）

Fig. 5　Slot filling (predicting slot label for “transformer”)

引擎Elasticsearch查询语句，返回符合的KG三元组或者Q/A对结果。

2.2.4　答案排序层

问题答案分为KG三元组+Elasticsearch文档、Q/A对+Elasticsearch文档2种组合，优先显示KG三元组或Q/A对结果，然后显示Elasticsearch文档。

3　设备知识图谱技术组件框架

设备知识图谱技术组件框架如图6所示。框架分为数据层、设备知识提取组件、设备图谱构建组件、设备知识服务组件4层。

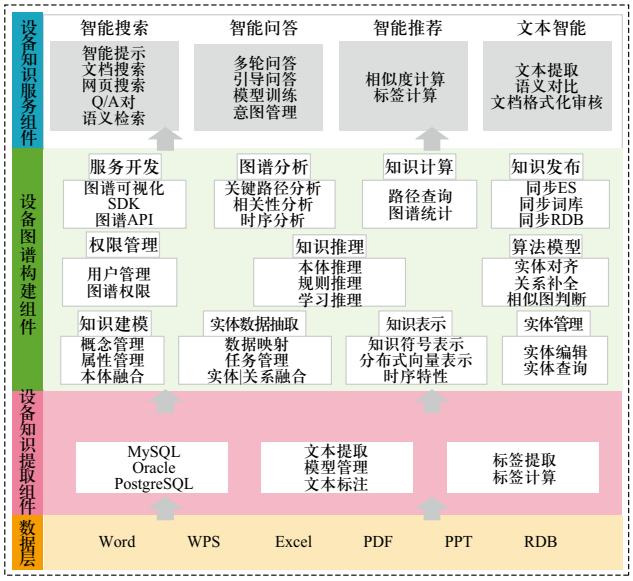


图 6　设备知识图谱技术组件框架

Fig. 6　Component framework of equipment knowledge graph

1）数据层：涵盖变压器各类数据源，包括RDB数据、长文本文档数据（如WPS、PDF等）。

2）设备知识提取组件：从数据层提供的数据中抽取变压器设备的各类知识，分为结构化知识抽取、非结构化知识抽取。

3）设备图谱构建组件：提供变压器本体定义、变压器实体知识存储、知识融合与推理、路径查询等服务。

4）设备知识服务组件：包含基于知识库的智能搜索、智能问答、文本智能等服务组件。

3.1　设备知识提取组件

变压器文本报告自动化提取过程如图7所示。

本文提出了针对变压器长文本数据进行知识提取的技术框架，框架提供变压器语料标注、提取模型训练、知识提取等功能，支持以流程化方式对变压器长文本报告知识的自动化提取。

图 7　变压器长文本报告抽取过程

Fig. 7　Extraction process of long text of transformer report

3.1.1　数据预处理

对变压器相关的报告进行格式转换、去除图片、去除目录等预处理操作。

3.1.2　本体结构设计

本文研究针对的是变压器设备领域，对本体结构的准确度要求很高，自上而下借助业务专家确定变压器本体结构，明确变压器相关概念的属性和关系。

3.1.3　语料标注

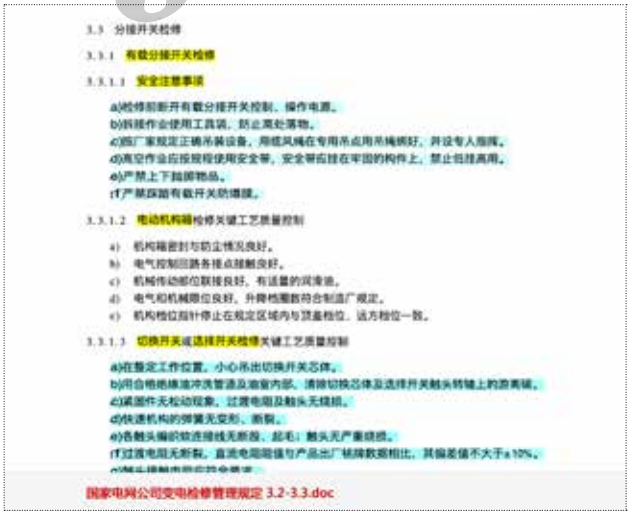
业务专家在语料标注工具上对文本进行标注，对语料库标注采用{B,I,O}（以变压器实体为例，B表示变压器实体的第一个词，I表示变压器实体的其余词，O表示不属于变压器实体的词）格式表示，基

于标注完成的语料库生成BERT-BiLSTM-CRF抽取模型。

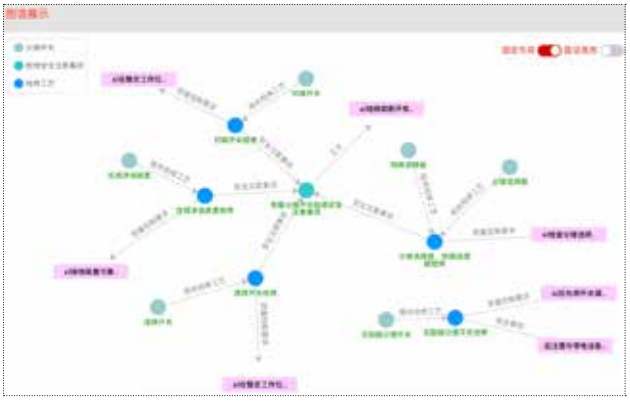
3.1.4 文档批量抽取

基于智能抽取模型，可对长文本报告中实体和关系进行批量自动识别，将大段非结构化文本转成结构化的图谱。以《国家电网公司变电检修管理规定》3.2—3.3节为例，原文部分内容如图8（a）所示，基于智能抽取模型识别后，将“有载分接开关检修”识别为检修工艺，将“a）检修前断开……f）严禁踩踏有载开关防爆膜”识别为“有载分接开关检修”实体的“安全注意事项”属性值。识别后形成的图谱如图8（b）所示。

图8（b）中，左上角图例不同颜色表示不同类型的实体，连线上的文字表示属性名称，粉色方框表示具体的属性值。



（a）长文本文档原文



（b）文档自动识别后的知识图谱

图 8 长文本报告原文及经过命名实体识别后形成的图谱

Fig. 8 Original text of document and knowledge graph formed by named entity recognition model

3.2 设备图谱构建组件

本文采用语义网网络本体语言（Ontology Web Language, OWL）标准作为知识表示模型，选择开源Jena TDB组件存储图谱本体概念模型，利用MongoDB数据库存储实体和关系三元组。通过对MongoDB性能增强实现支持对三元组进行大规模存储的语义数据库组件，提供对变压器图谱的知识存储、融合、推理等服务。

3.3 设备知识服务组件

设备知识服务组件包含了基于知识库的智能搜索、智能问答等。

3.3.1 智能搜索组件

基于KG的意图识别和槽位提取模型，提供了语义联想、知识推理、智能排序、意图推理等能力，判断用户查询意图，提取用户输入槽位，精准定位用户查询的实体，帮助用户更准确搜索到目标。

3.3.2 智能问答组件

支持6种类型的问句，包括单实体、单实体属性、单实体关系、近义问句、集合、带算子的问句，并通过问答上下文分析引擎，补全对问句的实体或属性（关系），从而支持以多轮方式进行灵活交互查询。6种类型问句如表1所示。

表 1 典型交互输入问句

Table 1 Typical questions

编号	典型问句	问句模板
类型1	四平变电站	{变电站实体}
类型2	西湖变电站1号变压器	{变电站实体}{变压器实体}
类型3	仙人桥变电站1号变压器生产厂家是哪里/哪里制造的仙人桥变电站1号变压器？	{变电站实体}{变压器实体}的{生产厂家，变压器属性}是哪里/{哪里制造，生产厂家}的{变电站实体}{变压器实体}
类型4	仙人桥变电站有多少变压器？	{变电站实体}有多少变压器
类型5	杭州电力公司XJ生产的存在硅胶变色缺陷的变压器都有哪些？	{地市公司实体}{XJ，生产厂家实体}存在{缺陷实体}的变压器有哪些？
类型6	仙人桥变电站有多少220 kV的变压器？	{变电站实体}有多少{220 kV，电压等级实体}的变压器

4 算例分析

本文对面向变压器智能运检的知识图谱构建和智

能问答技术进行了算例验证与分析，包括变压器实体抽取模型、知识融合和存储模型、变压器实体信息智能问答模型。

4.1　变压器实体抽取模型

4.1.1　模型准备

按本文2.1.2节所描述的非结构化文本数据知识抽取技术方法，分别构建了CRF、BiLSTM、BERT-BiLSTM-CRF三个模型进行了实验。模型基于Google开发的端到端开源深度学习框架平台Tensorflow搭建并确定模型的超参数。在选择超参数时，使用默认的超参数设置，观察损失函数值（loss）的变化，初步确定各个超参数的范围，再进行调参。对于每个超参数，在每次调整时，只调整一个参数，然后观察loss值变化，其中，训练集批次大小参数（batch_size）的选择参考GPU缓存大小，丢弃参数（dropout）取默认值，序列长度参数（seq_length）根据语料库的句子长度确定，以能覆盖大部分句子为准，最终确定的实验参数如下：seq_length为128，lstm单元数参数（lstm_size）为128，训练集批次大小参数（batch_size）为64，测试集批次大小参数（batch_size）为64，学习率参数为10^{-5}。为防止训练中出现梯度爆炸，使用梯度裁减技术并设置裁剪参数（clip）为0.5，dropout值为0.5。

4.1.2　实验数据及验证指标

收集历年来设备领域与变压器故障报告，对报告进行预处理后，形成训练集与测试集。

1）变压器语料收集，来源包括故障报告、设备状态评价报告、家族性缺陷报告等。对语料数据进行预处理，保留文本数据，去除报告中的图片、表格、目录、图表等元素，并对文本进行分段、分句处理，形成可独立标注的典型语句。

2）对语料库进行切分，选择变压器故障案例报告298份，涉及变压器概念共34个，其中核心概念如表2所示，将报告按照一定比例切分，232份用于训练，66份用于测试。

表 2　故障报告实体类型及属性列表

Table 2　Entity type and attribute list of fault reports

编号	实体类型	属性列表	数量
1	变压器	名称、电压等级、型号等	228
2	故障	故障时间、运行状态、跳闸状态等	76
3	生产厂家	厂家名称	243
4	设备部件	出厂时间、型号、投运日期	35
5	故障原因	故障原因、责任原因	237
6	试验结论	是否合格	190
7	检查结论	结论内容	319
8	油色谱试验	试验时间、试验结论	22
9	变电站	变电站名称、电压等级	217
10	故障现象	现象标签	193

3）语料库序列标注，对语料库训练集中的报告进行序列标注，依据变压器故障概念，对语料库处理后的数据以句子为单位进行标注。本文使用三元标记集{B,I,O}。测试集中主要实体的数量如表3所示。

表 3　测试集故障报告实体数量

Table 3　Statistics of entities in test set

实体类型	实体数量	实体类型	实体数量
变压器	62	故障原因	67
故障	15	试验结论	51
生产厂家	71	检查结论	84
设备部件	11	油色谱试验	29
故障现象	54	变电站	60

对各类命名实体，采用P、R及F_1值作为模型评价指标，定义如式（5）—（7）。

$$P=\frac{\text{正确识别出的命名实体个数}}{\text{识别出的命名实体个数}}\times 100\% \tag{5}$$

$$R=\frac{\text{正确识别出的命名实体个数}}{\text{标准结果中命名实体个数}}\times 100\% \tag{6}$$

$$F_1=\frac{2\times P\times R}{P+R}\times 100\% \tag{7}$$

4.1.3　实验结果

在数据集上，采用CRF、BiLSTM、BERT-BiLSTM-CRF模型进行分析。实验结果如表4所示，可看出，基于BERT-BiLSTM-CRF的模型比CRF、LSTM模型准确率高。

表 4　在数据集上应用各模型的指标对比

Table 4　Comparison of various models index

%

任务	R	P	F_1
CRF	61.30	65.13	63.16
LSTM	63.45	66.78	65.07
BERT-BiLSTM-CRF	72.37	76.51	74.38

针对各实体类型，提取模型的实验结果如表5所示。

表 5 各类实体的识别结果对比

Table 5 Comparison of entities recognition

%

任务	R	P	F_1
变压器	84.50	81.37	82.90
故障	94.90	92.08	93.47
生产厂家	89.18	87.82	88.49
设备部件	69.57	66.67	68.09
故障原因	72.31	71.26	71.78
试验结论	10.00	9.09	9.52
检查结论	75.00	75.00	75.00
油色谱试验	77.78	77.78	77.78
变电站	81.13	78.03	79.55
故障现象	73.81	68.89	71.26

4.2 知识融合和存储模型

4.2.1 知识融合模型

本文的实体融合模型根据2.1.3节的实体特征量相似度计算公式实现。以从变压器故障报告中提取的实体进行知识融合实验，选取了变电站、变压器、生产厂家、地市公司、故障现象共5类实体，5类实体对应的概念间关系如图9所示，图谱中各类实体数量如表6所示。

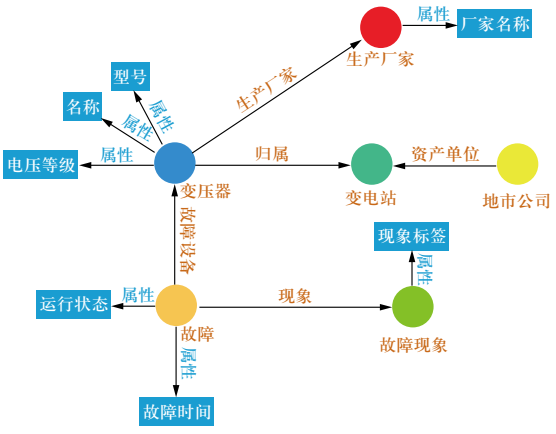


图 9 变压器知识图谱结构

Fig. 9 Schema of transformer knowledge graph

对于细分领域的知识化融合场景，当从文档中提取新的实体，将实体存储到图谱库时，需要准确找到与新增的实体存在冲突的实体，因此，选择P指标作为融合模型的评价指标。

表 6 变压器故障图谱中实体数量

Table 6 Statistics of entities in fault knowledge graph

实体类型	变电站	变压器	生产厂家	地市公司	故障现象
数量	217	228	243	10	193

选择31份变压器故障报告作为测试集，测试集包含了变电站、变压器、生产厂家、地市公司、故障现象5类实体，且提取后的各实体之间建立了关系。各类实体的分布及通过融合模型识别出的实体数量分布如表7所示。

表 7 测试集实体数量及模型识别结果

Table 7 Statistics of entities in test set and the precision of entity fusing model

实体类型	测试集实体数量	识别正确数量	准确率/%
变电站	25	22	88.00
变压器	31	27	87.10
生产厂家	48	35	72.92
地市公司	25	24	96.00
故障现象	74	73	98.65

从表7的数据可以看出，对故障现象、地市公司的实体融合模型准确率较高，主要是因为故障现象、地市公司实体在抽取实体时进行了归一化处理，而生产厂家、变压器由于存在简称、关系不完善、属性值缺失等特征，导致模型的准确率相对较低。

4.2.2 知识存储模型

本文基于MongoDB的文档结构设计了一种存储语义三元组数据的存储模型，覆盖RDF的SPO三元组，模型采用轻量级的JavaScript数据交换格式（JavaScript Object Notation, JSON）通过5类字段进行定义，各字段定义如表8所示。

表 8 实体存储模型字段

Table 8 Columns of entities storage model

编号	三元组元素	属性
1	主语（S）	ID、业务ID、对应的概念
2	谓语（P）	ID、英文名称、属性
3	宾语（O）	ID、业务ID、对应的概念、值
4	其他字段	开始时间、结束时间
5	预留字段	用于后续扩充

基于对SPO三元素的7层组合索引，包括S、P、O、SP、PO、SO、SPO，该存储模型能够提供对数据的高效查询。

4.3　变压器实体信息智能问答模型

4.3.1　模型准备

本文的变压器实体信息智能问答模型中，意图识别及槽位提取模型基于Tensorflow框架搭建，采用本文2.2.2节描述的算法模型实现，batch_size参数值为32，词向量维度（words_embedding_dim）参数值为300，dropout值为0，训练轮次（epoch_num）参数值为20，训练形成的意图识别模型和槽位识别模型采用BiLSTM模型结构，该模型包含2层的LSTM模型，LSTM层节点数是50，输入维度是300，隐层维度值是300，具体结构参见图4与图5。

4.3.2　实验数据及验证指标

本文以变压器故障报告进行实验，选取了变电站、变压器等共6类概念，各类实体数量如表9，6类问题模板如表1。根据6类问句模板进行问句生成，共生成问句198 100条，每类问句最多5万条，最后按照8∶2的比例，将报告数据分为训练集和验证集。训练集与验证集6类语句的分布如表10所示。

表 9　实验数据概念列表及数量

Table 9　Statistics of concepts in experimental data

实体类型	变电站	变压器	生产厂家	地市公司	电压等级	缺陷
数量	100	240	60	12	3	5000

表 10　六类语句训练集和验证集的数量

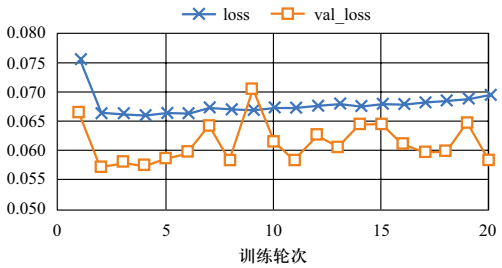
Table 10　Statistics of six type sentences in training sets and validation sets

问句类型	类型1	类型2	类型3	类型4	类型5	类型6
训练集数量	80	19 200	40 000	19 200	40 000	40 000
验证集数量	20	4800	10 000	4800	10 000	10 000

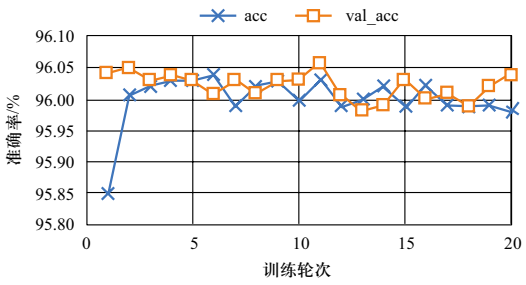
模型的评价指标包括模型在训练集上的损失值（loss）、在训练集上模型的准确率（acc）、模型在验证集上的loss值（val_loss）、模型在验证集上模型的准确率（val_acc）4个指标，通过多轮次训练形成的模型进行自动筛选，以val_loss最小来选取最优模型。

4.3.3　实验结果

意图识别模型的loss、val_loss、acc、val_acc等指标与训练轮次的关系如图10所示。



（a）意图识别模型loss与val_loss指标

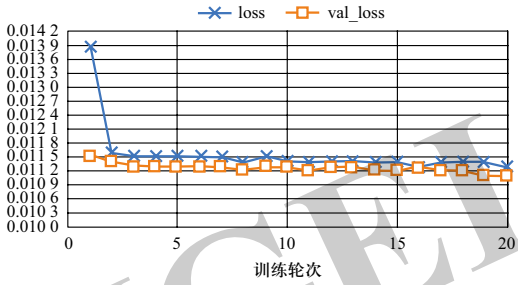


（b）意图识别模型acc与val_acc指标

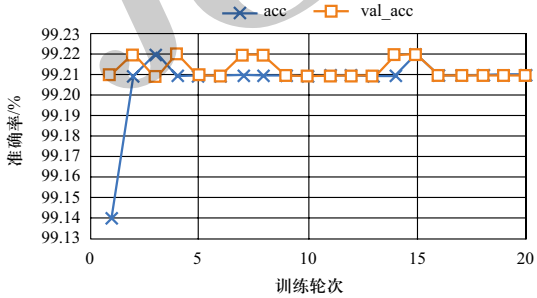
图 10　意图识别模型评价指标变化曲线

Fig. 10　Curve of intent recognition model evaluation index value changes

槽位提取模型的loss、val_loss、acc、val_acc等指标与训练轮次的关系如图11所示。



（a）槽位提取模型loss与val_loss指标



（b）槽位提取模型acc与val_acc指标

图 11　槽位提取模型评价指标变化曲线

Fig. 11　Curve of slot model evaluation index value changes

5 应用成果

本文研究成果在某省电力公司电力主设备知识库系统、某电科院变压器设备知识服务平台系统中得到应用，为变压器故障报告提取、变压器设备信息灵活问答等场景实现智能化提升。

5.1 变压器故障报告提取

本文研究的变压器实体抽取模型在某电科院变压器知识服务平台中进行了应用，对近十年来各电压等级变压器设备故障报告进行了自动化提取，每份报告（平均7~10页）平均提取时间约2 s，变压器核心实体（变压器、生产厂家、故障等）识别准确率大于80%。提取结果示例如图12所示。

图 12 变压器故障报告提取示例

Fig. 12 Sample of transformer fault report extraction

5.2 变压器设备信息灵活问答

本文基于BiLSTM实现的变压器实体智能问答模型解决了传统浅层学习算法对数据挖掘能力有限的问题[18]，实现了对所管辖变压器台账与故障信息的灵活查询，通过多轮交互式问答，为检修公司一线员工提供了设备台账信息、缺陷信息、标准导则等的灵活查询，典型问句回答准确率在90%以上。问答示例如图13所示。

6 结语

本文总结了近年来作者在变压器设备领域应用知识图谱技术的研究与实践，研究了基于BERT-BiLSTM-CRF的变压器实体抽取模型、意图识别与槽位提取模型等算法模型，提出了设备知识图谱技术组件框架，并将该框架在电力公司的变压器设备信息灵活问答、变压器故障报告自动化提取场景中进行了验证，受篇幅限制，部分模型及算法技术细节未进行深入说明。本文研究成果不局限于变压器设备，也适用于断路器等其它输变电设备，具有较强的推广应用价值。

图 13 变压器智能问答示例

Fig. 13 Sample of transformer question answering

在应用实践中，目前知识图谱技术在设备智能运检领域的工业化成熟度还需进一步提升。电力行业属于重资产、高安全行业，要求知识图谱技术在小样本学习、自训练等场景下有良好表现，为知识图谱技术在设备智能运检领域的更深入应用指出了方向。

参考文献

[1] 国网山东省电力公司. 人工智能平台白皮书[R]. 济南：国网山东省电力公司，2019.

[2] 中国电机工程学会电力信息化专委会. 中国电力大数据发展白皮书[R]. 北京：中国电机工程学会，2013.

[3] 邱剑. 电力中文文本数据挖掘技术及其在可靠性中的应用研究[D]. 杭州：浙江大学，2016.

[4] 施萱轩，姜红红，梁浩，等. 文本挖掘技术研究及其在电力行业的应用[J]. 机电信息，2017(30)：42-45.

[5] 张鹏，王玮，赵德伟，等. 基于文本挖掘的电力设备缺陷用户画像构建[J]. 科技风，2019(33)：177-180.

[6] 杨兵，丁辉，罗为民，等. 基于知识库的变压器故障诊断专家系统[J]. 中国电机工程学报，2002，22(10)：121-124.
YANG Bing, DING Hui, LUO Weimin, et al. Expert system of transformer fault diagnosis based on knowledge base[J]. Proceedings of the CSEE, 2002, 22(10)：121-124(in Chinese).

[7] 刘梓权，王慧芳. 基于知识图谱技术的电力设备缺陷记录检索方法[J]. 电力系统自动化，2018，42(14)：158-164.
LIU Ziquan, WANG Huifang. Retrieval method for defect records of power equipment based on knowledge graph technology[J]. Automation of Electric Power Systems, 2018, 42(14): 158-164(in Chinese).

[8] 王璟. 电力变压器故障诊断策略分析与设计[D]. 济南：山东大学，2016.

[9] 姬源，谢冬，周思明，等. 电力领域语义搜索系统的构建方法[J]. 计算机系统应用，2016，25(4)：91-96.
JI Yuan, XIE Dong, ZHOU Siming, et al. Construction method of semantic search system in power domain[J]. Computer Systems & Applications, 2016, 25(4): 91-96(in Chinese).

[10] 田晓，刘勇超，王婧，等. 电网公司客户服务知识图谱构建的应用价值[J]. 山东电力技术，2015，42(12)：65-67.
TIAN Xiao, LIU Yongchao, WANG Jing, et al. Application value of building knowledge graph system for hotline customer service in state grid corporation[J]. Shandong Electric Power, 2015, 42(12): 65-67(in Chinese).

[11] 张淼. 基于中文知识图谱的智能问答系统设计与实现[D]. 武汉：华中师范大学，2018.

[12] MANNING C D, RAGHAVAN P. An introduction to information retrieval[M]. Cambridge: Cambridge University Press, 2009.

[13] MANNING C D. Computational linguistics and deep learning[J]. Computational Linguistics, 2015, 41(4): 701-707.

[14] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. 2018: arXiv:1810.04805[cs.CL]. https://arxiv.org/abs/1810.04805.

[15] 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.
XU Zenglin, SHENG Yongpan, HE Lirong, et al. Review on knowledge graph techniques[J]. Journal of University of Electronic Science and Technology of China, 2016, 45(4): 589-606(in Chinese).

[16] 王子牛，姜猛，高建瓴，等. 基于BERT的中文命名实体识别方法[J]. 计算机科学，2019，46(增刊2)：138-142.
WANG Ziniu, JIANG Meng, GAO Jianling, et al. Chinese named entity recognition method based on BERT[J]. Computer Science, 2019, 46(Supplement 2): 138-142(in Chinese).

[17] 王鑫，邹磊，王朝坤，等. 知识图谱数据管理研究综述[J]. 软件学报，2019，30(7)：2139-2174.
WANG Xin, ZOU Lei, WANG Chaokun, et al. Research on knowledge graph data management: a survey[J]. Journal of Software, 2019, 30(7): 2139-2174(in Chinese).

[18] 蒋逸雯，李黎，李智威，等. 基于深度语义学习的电力变压器运维文本信息挖掘方法[J]. 中国电机工程学报，2019，39(14)：4162-4171.
JIANG Yiwen, LI Li, LI Zhiwei, et al. An information mining method of power transformer operation and maintenance texts based on deep semantic learning[J]. Proceedings of the CSEE, 2019, 39(14): 4162-4171(in Chinese).

[19] XU Kun, LAI Yuxuan, FENG Yansong, et al. Enhancing key-value memory neural networks for knowledge based question answering[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. June 2019, Minneapolis, Minnesota, 2019: 2937-2947.

[20] 周博通，孙承杰，林磊，等. 基于LSTM的大规模知识库自动问答[J]. 北京大学学报（自然科学版），2018，54(2)：286-292.
ZHOU Botong, SUN Chengjie, LIN Lei, et al. LSTM based question answering for large scale knowledge base[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2018, 54(2): 286-292(in Chinese).

[21] Bingfeng Luo, Yansong Feng, Zheng Wang, et al. Marrying up regular expressions with neural networks: a case study for spoken language understanding[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. April, 2018, Melbourne, Australia, 2018: 2083-2093.

[22] LAI Y X, JIA Y Y, LIN Y, et al. A Chinese question answering system for single-relation factoid questions[C]//Natural Language Processing and Chinese Computing. Cham: Springer International Publishing, 2018: 124-135.

收稿日期：2020-07-09；修回日期：2020-08-21。

作者简介：

张敏杰

张敏杰（1983），男，硕士，研究方向为电力企业人工智能技术，E-mail：zhangmj@kinditec.com。

徐宁（1993），男，硕士，研究方向为电力电子技术在电力系统中的应用，E-mail：ningxu@zju.edu.cn。

王宇飞（1981），男，硕士，研究方向为认知智能技术。通信作者，E-mail：wangyf@kinditec.com。

（责任编辑　李锡）