

基于 MSSI-UIE 的电力调度事件抽取方法研究

谭琦¹, 朱道坤¹, 唐昊¹, 朱克东², 杨慧彪³

- (1. 合肥工业大学电气与自动化工程学院, 安徽省 合肥市 230009;
2. 中国电力科学研究院有限公司(南京), 江苏省 南京市 210003;
3. 国网宁夏电力有限公司电力科学研究院, 宁夏回族自治区 银川市 750001)

Power Dispatching Event Extraction Method Based on MSSI-UIE

TAN Qi¹, ZHU Daokun¹, TANG Hao¹, ZHU Kedong², YANG Huibiao³

- (1. School of Electrical Engineering and Automation, Hefei University of Technology, Hefei 230009, Anhui Province, China;
2. China Electric Power Research Institute (Nanjing), Nanjing 210003, Jiangsu Province, China; 3. Electric Power Research Institute of State Grid Ningxia Electric Power Co., Ltd., Yinchuan 750001, Ningxia Hui Autonomous Region, China)

ABSTRACT: The power dispatching event knowledge graph can provide comprehensive historical dispatching knowledge to assist dispatchers in decision-making. However, the complex logical structure of power dispatching texts, combined with data sparsity and few-shot scenarios, makes event argument extraction challenging, which severely hampers the construction of knowledge graphs. To address these issues, this paper proposes a novel multiple structural schema instructor- universal information extraction (MSSI-UIE) framework for power dispatching event extraction. First, the UIE's original generative architecture is improved with the bidirectional encoder representations from transformers (BERT) model, integrating power domain-specific term masking pre-training and dual-pointer network to achieve precise event argument localization. Then, a DeepSeek-based text data augmentation approach is developed to effectively expand the training corpus. Finally, three types of structural schema instructor (SSI) are designed to collaboratively extract event arguments from multiple semantic dimensions. Experimental results demonstrate that this method exhibits strong performance in data-sparse and few-shot scenarios, effectively enhancing the efficiency and accuracy of power dispatching event extraction.

KEY WORDS: knowledge graph; power dispatching event; event extraction; multiple structural schema instructor; universal information extraction

摘要: 电力调度事件知识图谱可以提供丰富的历史调度知

识, 辅助调度员进行调度决策。然而, 电力调度文本逻辑结构复杂, 在数据稀疏和小样本场景下事件论元抽取困难, 严重制约了知识图谱的构建。针对上述问题, 该文提出一种基于多种结构化模式提示器与通用信息抽取的电力调度事件抽取方法。首先, 采用基于 Transformer 的双向编码器表示 (bidirectional encoder representations from transformers, BERT) 模型改进通用信息抽取 (universal information extraction, UIE) 原有的生成式架构, 结合电力实词掩码预训练和双指针网络, 实现事件论元的精准定位; 然后, 设计基于 DeepSeek 的调度事件文本数据增强方法, 实现训练数据的有效扩增; 最后, 设计 3 种结构化模式提示器, 从多个语义维度对事件论元进行联合抽取。实验结果表明, 在数据稀疏以及小样本场景下, 该方法能够有效提升电力调度事件抽取的效率与准确性。

关键词: 知识图谱; 电力调度事件; 事件抽取; 多结构化模式提示器; 通用信息抽取

0 引言

随着新型电力系统建设持续推进、设备规模持续增长, 调度经验和电网物理结构知识日益丰富, 调度决策难度也随之不断增加。传统电力调度模式由于未对调度知识进行规范化建模与管理, 调度员决策主要依赖个人经验, 并需查阅规程和预案等文件, 导致效率低下、易出疏漏^[1]。同时, 调度员在进行调度决策时积累的大量经验知识难以有效分享和利用, 依靠经验积累和认知水平的个体决策已难以满足当下电网调度决策的需求^[2]。知识图谱作为结构化的语义知识库, 通过符号形式描述物理世界中的实体及其相互关系, 可实现电力调度知识的

基金项目: 国家电网有限公司科技项目(5108-202335450A-3-2-ZN)。

Project Supported by Science and Technology Project of State Grid Corporation (5108-202335450A-3-2-ZN).

规范化管理^[3]。在此基础上,事件知识图谱通过提取文本中的事件论元(如主体、时间、地点等),进一步实现事件知识的存储与利用^[4]。面向电力系统调度日志中存在的海量调度经验知识,研究事件论元抽取方法,构建更符合电力专家和从业人员认知习惯的电力调度事件知识图谱,可以为调度员决策提供知识支撑,具有重要的现实意义^[5]。

事件图谱的构建涵盖数据获取、事件抽取^[6]、知识融合等核心步骤,其中事件抽取作为关键环节,直接影响图谱的质量与应用价值。电力调度文本的事件抽取,不仅要求兼顾文字的字面含义,还要求结合领域背景和专业术语等要素。传统事件抽取方法采用 BIO 和 BIOES 等方式标注事件文本^[7],利用标注数据训练长短期记忆(long short-term memory, LSTM)神经网络,以实现事件触发词^[8]和事件论元^[9]的抽取。然而,这类事件抽取方法依赖大量标注数据,如文献[10-11]在电力领域进行实体关系抽取时,需要通过大量人工标注数据的训练,才使抽取方法具有良好的抽取能力。近年来,考虑到数据标注巨大的时间和人工成本,通用领域中的小样本学习受到研究人员的广泛关注,其实现方法包括数据增强、元学习、迁移学习等^[12]。文献[13]将基于 Transformer 的双向编码器(bidirectional encoder representations from transformers, BERT)预训练模型与 LSTM 神经网络相结合,将通用领域的语义知识迁移至电力领域;文献[14]采用同类实体替换的方法扩充训练数据集,以提升小样本场景下模型的抽取能力,但该方法可能改变原始语义,进而影响模型训练效果;文献[15]将事件抽取转化为问答形式,基于机器阅读理解(machine reading comprehension, MRC)框架,通过提问实现事件抽取,但该方法会导致误差传播,降低模型抽取能力;文献[16]为了避免误差传播的产生,使用独立的模型来抽取事件触发词,但增加了额外的训练成本;文献[17]提出的通用信息抽取(universal information extraction, UIE)方法利用结构化模式提示器(structural schema instructor, SSI)缓解了误差传播的问题,但其生成式架构导致模型训练成本较高。上述方法虽然在通用领域具有较好的抽取效果,但面对专业性较高和句式复杂多变的电力调度文本时,难以实现理想的抽取效果。

综上,为实现电力调度文本的自动和准确抽

取,以充分利用调度日志中的经验知识,本文提出一种基于多种结构化模式提示器与通用信息抽取(multiple structural schema instructor-universal information extraction, MSSI-UIE)的电力调度事件抽取方法。首先,采用掩码语言模型改进了 UIE 原有的生成式架构,通过对电力专业术语进行实词掩码预训练,提高模型在电力领域的语义理解能力;其次,提出基于 DeepSeek 的数据增强方法,提升模型对稀疏和小样本事件的抽取能力;最后,提出 MSSI 联合抽取方法,通过设计多种 SSI 对事件论元进行联合抽取,提高事件抽取的可靠性。

1 电力调度文本事件抽取特点

电力调度日志涉及故障处置、调度决策、设备检修等多维度知识,对电力调度文本进行事件抽取的难点总结如下:

1) 电力调度文本专业性较高,具有大量的专业名词。例如“***四光 35 kV***线非停”,这个事件内就包含了“***四光”、“35 kV”、“***线”和“非停”4个专有名词,通用领域模型的抽取效果不佳。

2) 我国电力系统具有良好的稳定性,相同故障调度事件的频次往往较低,用于模型训练时,存在部分事件类型文本数据量不足、数据不平衡等数据稀疏性问题,难以有效学习这类事件的语句特征。

3) 电力调度文本具有事件类型丰富、句式复杂多变、不同事件类型间语义结构差异大等特点,这对算法的事件抽取能力提出了较高要求。

本文在设计事件抽取方法时,充分考虑了上述调度文本特点,基于电力实词掩码预训练的 BERT 对电力调度文本进行语义向量编码,并针对数据稀疏以及小样本等问题,设计基于大模型的数据增强方法与 MSSI 事件联合抽取方法,提高事件抽取的可靠性。

2 基于 MSSI-UIE 的电力调度事件抽取方法

基于 MSSI-UIE 的电力调度事件抽取方法框架如图 1 所示,包含数据处理、数据增强和 UIE3 个核心模块。首先,将原始调度日志文本数据与数据增强模块生成电力调度文本处理成 MSSI、TXT 和 Result 共 3 部分, MSSI 是由 1 个 Schema 构造出的多种 SSI, TXT 是待抽取文本, Result 是标注结果;然后,将“SSI+TXT”送入 BERT 编码器,并在编

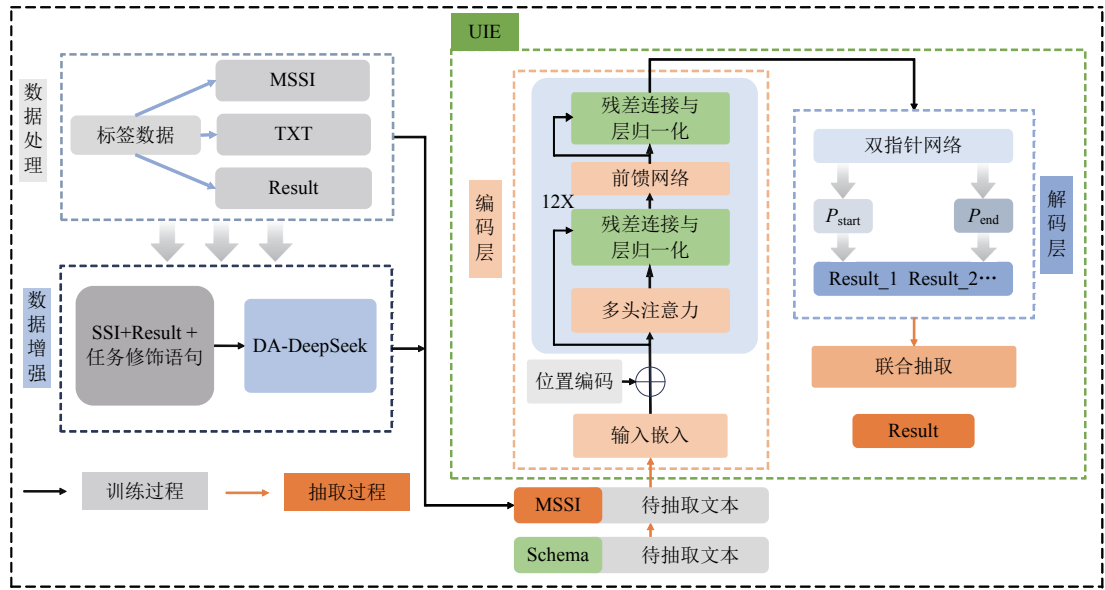


图 1 基于 MSSI-UIE 的电力调度事件抽取模型

Fig. 1 Power grid dispatch event extraction model based on MSSI-UIE

码器最后一层接入双指针网络，负责解码事件触发词和事件论元；最后，提出 MSSI 联合抽取方法对解码得到的事件论元进行整合。

2.1 基于预训练模型的抽取任务转化与字向量编码

2.1.1 基于编码器的事件抽取任务转化

事件是由多个事件三元组(事件触发词、关系、事件论元)构成的。本文结合 UIE 中的 SSI、Schema、Prompt Learning 以及 MRC 的思想^[18]，提出将事件触发词、事件论元以及事件三元组抽取转化为基于 SSI 的抽取模式，并采用掩码语言模型与双指针解码框架替代 UIE 框架中原有的生成式框架。依据事件抽取 Schema 转化成的 SSI^[17]，抽取事件触发词和事件论元，从而获取事件三元组。

本文采用 UIE 模型，基于 Prompt Learning 思想，将 SSI 作为 MRC 机制中的问题，与调度文本 TXT 进行拼接，将拼接结果输入模型，从而输出事件论元和事件三元组：

$$Y = \text{UIE}(S \oplus T)$$
 (1)

式中： S 为事件抽取 SSI； T 为待抽取文本； Y 为抽取结果； $\oplus$ 为文本拼接符号。

如图 2 所示，事件论元的抽取采用一步式方法，即将“事件论元 SSI+TXT”拼接后输入模型，直接得到论元结果。而事件三元组的抽取则分为两步：1) 将“事件触发词标签 SSI+TXT”输入模型，获取事件触发词；2) 将触发词与对应的关系类型组合为新的“事件三元组 SSI+TXT”，再次输入模型，最终输出包含相应论元的事件三元组。

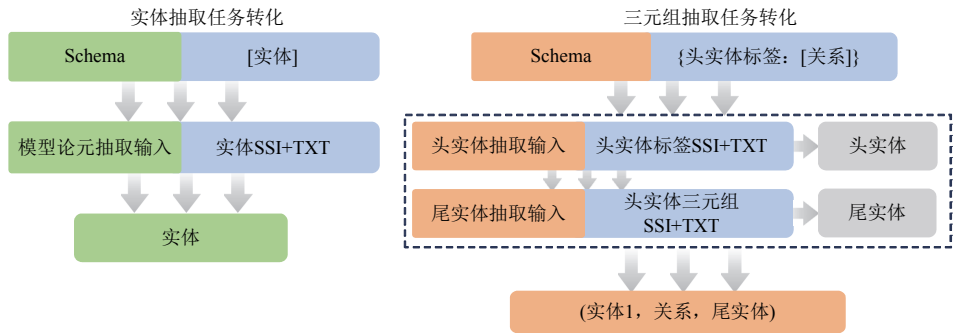


图 2 事件抽取任务转化形式

Fig. 2 Transformation forms of event extraction tasks

2.1.2 编码器掩码策略

掩码预测是掩码类编码器的预训练任务之一，通过自监督训练让模型理解文本中的语义信息。BERT 和 ERNIE 是目前主流的掩码类编码器，两者

的掩码策略如图 3 所示，其中 BERT 是在文本中随机挑选 15% 的字符进行静态掩码操作^[19]，以学习单个字符的上下文语义信息；ERNIE 在 BERT 的基础上针对中文文本增加了多种掩码策略，对实词和短

句进行掩码操作，以学习在词法和句法层面的上下文关联及语义信息，提高中文文本的事件抽取效果。

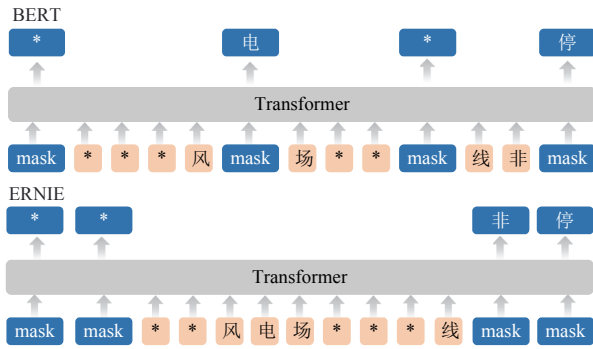


图 3 BERT 与 ERNIE 掩码策略

Fig. 3 BERT and ERNIE's mask strategy

2.1.3 输入表示与编码

在编码层，给定一段电力调度文本后，掩码模型会将 SSI 和待抽取文本中的每个字符转化为包含词嵌入和位置编码的高维词向量，其中词嵌入用于表示字符的语义信息，位置编码用于记录当前字符的绝对位置信息，从而得到模型输入的高维词向量序列：

$$S \oplus T = [\text{cls}, s_1, s_2, \dots, s_n, [\text{sep}], x_1, x_2, \dots, x_m] \quad (2)$$

式中：[cls]、[sep]为掩码语言模型的特殊字符； $s_1, s_2, \dots, s_n$ 为 SSI 的高维词向量； $x_1, x_2, \dots, x_m$ 为文本内容 T 的高维词向量。

随后将 $S \oplus T$ 送入 Transformer 模块通过自注意力机制进行编码，其注意力机制的计算如下：

$$Q, K, V = W_q \cdot (S \oplus T), W_k \cdot (S \oplus T), W_v \cdot (S \oplus T) \quad (3)$$

式中： Q 、 K 、 V 分别为查询矩阵、键矩阵、值矩阵； W_q 、 W_k 、 W_v 为权重矩阵。

在计算注意力权重时， Q 和 K 通过点积运算得到字符间的相关性分数，为了消除文本长度对数值大小的影响，将点积结果除以 $\sqrt{d}$ 进行缩放。随后，通过 Softmax 函数进行归一化处理，得到注意力权重矩阵 A ：

$$A = \text{Softmax}(Q^T K / \sqrt{d}) \quad (4)$$

式中 d 为键矩阵的维度。

最终，将注意力权重 A 与值矩阵 V 相乘，可以得到电力调度语境化编码向量 X_A ：

$$X_A = \text{Attention}(Q, K, V) = A \cdot V \quad (5)$$

式中 $\text{Attention}(\cdot)$ 为注意力计算函数。

为了获得不同维度的语义信息，本文基于多头注意力机制进行多次计算并拼接结果，再使用全连接层得到多头注意力编码 X_m ：

$$X_m = W_m \cdot \text{Concat}[A \cdot V]^h \quad (6)$$

式中： $\text{Concat}[\cdot]$ 为矩阵拼接函数； h 为多头注意力的多头数； W_m 为全连接层的权重矩阵。

2.2 双指针解码器

本文以 2 个并行全连接层以及 Softmax 函数作为双指针解码器，预测当前实体在文中的起始和结束位置并根据阈值进行筛选。如图 4 所示，所预测实体为整个文本序列。同时，保留起始和结束指针的概率作为“可信度”为后续联合抽取提供参考。

文本序列	*	*	*	风	机	温	升	高
起始位置 概率	0.64	0.11	0.05	0.02	0.07	0.06	0.04	0.01
结束位置 概率	0.02	0.01	0.01	0.01	0.12	0.09	0.01	0.73

图 4 双指针解码过程

Fig. 4 Process of double pointer decoding

起始位置与结束位置的概率计算公式如下：

$$\begin{cases} P_{\text{start}} = \text{Softmax}(W_s \cdot X_m^i + b_s) \\ P_{\text{end}} = \text{Softmax}(W_e \cdot X_m^i + b_e) \end{cases} \quad (7)$$

式中： P_{start} 、 P_{end} 分别为实体起始与结束位置的概率分布向量； W_s 、 b_s 分别为起始线性层的权重矩阵和偏置向量； W_e 、 b_e 分别为结束线性层的权重矩阵和偏置向量； $\text{Softmax}(\cdot)$ 为分类函数； X_m^i 为 X_m 中第 i 个字符对应的词向量。

分别选取 P_{start} 与 P_{end} 中概率最大的位置作为预测的实体起始位置与结束位置，记作 M_{start} 与 M_{end} 。起始位置与结束位置的损失函数计算公式如下：

$$\begin{cases} L_{\text{start}} = \text{cross_entropy}(M_{\text{start}}, I_{\text{start}}) \\ L_{\text{end}} = \text{cross_entropy}(M_{\text{end}}, I_{\text{end}}) \\ L_{\text{loss}} = (L_{\text{start}} + L_{\text{end}}) / 2 \end{cases} \quad (8)$$

式中： $\text{cross_entropy}(\cdot)$ 为交叉熵损失函数； L_{start} 、 L_{end} 分别为实体起始位置与结束位置预测的损失值； I_{start} 、 I_{end} 分别为实际的实体起始位置与结束位置； L_{loss} 为起始指针和结束指针的平均损失值。

2.3 MSSI 联合抽取

为了提高事件抽取的准确性，本文设计一种基于 MSSI 的联合抽取方法，通过 MSSI 充分利用调度文本中的语义信息，从多个语义维度抽取事件论元，提高事件论元抽取的可靠性。如图 5 所示，该方法将传统的事件抽取 Schema 转化为 3 种具有递进语义利用层次的 SSI：1) 通用型事件论元 SSI 仅利用单个事件论元在调度日志中的局部语义信息；2) 触发词引导型事件论元 SSI 通过融合事件触发

词标签与论元的联合语义特征,增强论元抽取时的上下文感知能力;3)事件三元组 SSI 进一步利用完整事件三元组的全局语义约束,利用事件结构的整体性提高抽取的可靠性。MSSI 联合抽取方法通过这种层级化语义利用策略,从局部到全局逐步强化事件要素的语义表征,显著提升了复杂场景下事件抽取的准确性。

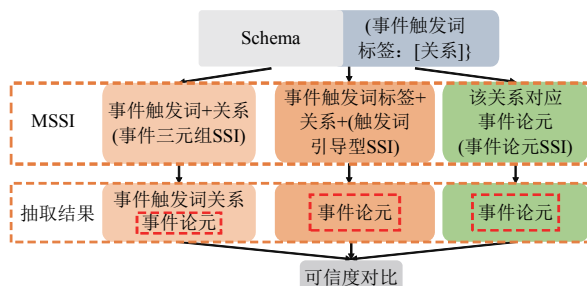


图5 Schema, MSSI 与抽取结果示意图

Fig. 5 Diagram of Schema, MSSI and extraction results

图6展示了本文提出的MSSI联合抽取的具体实现流程。例如:对“***风电场***线非停,***风机温升高,***线停电配合消缺。”进行事件原因的联合抽取。首先,本文通过自动转化机制,将事件抽取 Schema 分别转换为以下3种形式的SSI输入:1)通用型论元SSI,其对应的问题模板为“事件原因是?”;2)触发词引导型论元SSI,对应的问题模板为“事件触发词标签的事件原因是?”;3)事件三元组SSI,对应的问题模板为“***线非停的事件原因是?”。其中,事件三元组型SSI中的“***线非停”是UIE方法预先抽取的事件触发词。然后,基于协同抽取机制,抽取3个候选结果,通过置信度评估选择最优结果。最终输出可靠的事件三元组结构化表示(***线非停,事件原因,***风机温升高)。

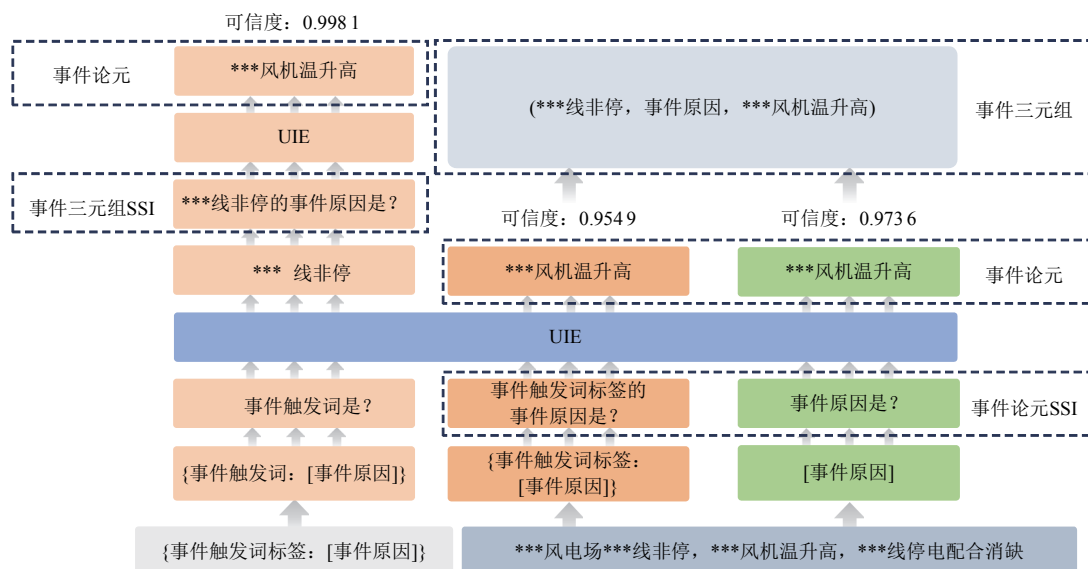


图6 MSSI联合抽取机制实现流程

Fig. 6 Workflow for implementing the MSSI joint extraction mechanism

2.4 数据增强

针对数据稀疏性的问题,本文提出一种基于大模型的数据增强方法,将SSI与Prompt Learning结合,使用SSI作为DeepSeek提示语句的重要组成部分,微调得到数据增强(data augmentation, DA)-DeepSeek模型,生成合理的电力调度文本TXT并转化成“SSI+TXT”的结构,为UIE模型提供更多高质量的训练语句。

如图7所示,将调度日志文本处理成SSI、Result、任务修饰语句和TXT,随后将“SSI+Result+任务修饰语句”作为DeepSeek数据增强的文本提示语句,“TXT”作为模型期望输出,并将文本提

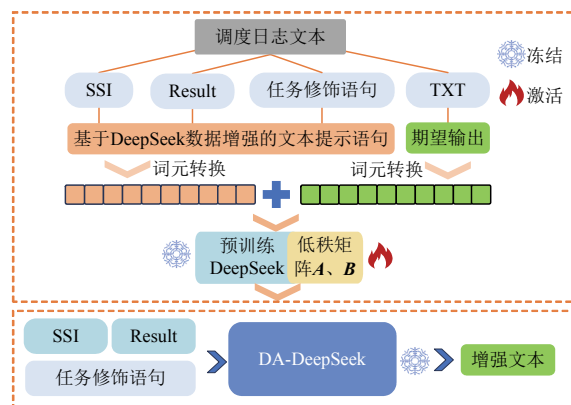


图7 基于DA-DeepSeek的数据增强流程

Fig. 7 Data augmentation process based on DA-DeepSeek
提示语句和期望输出分别进行词元(Token)转换,映射

为模型可处理的向量序列,供 DeepSeek 进行微调。同时,由于大模型训练成本高昂,本文采用 Lora 微调技术,冻结大模型原始参数,引入低秩矩阵 A 和 B ,在保证大模型性能的同时,减少训练参数量,使其可以生成合理的调度文本。

2.5 数据处理

2.5.1 事件标注

电力系统调度日志包含了故障信息、调度信息、检修信息等各种调度知识,文本句式复杂多变。为了提高 MSSI 联合抽取效果,本文采用事件论元与事件关系相同命名的标注方式,即事件触发词与事件论元之间的关系标签与事件论元标签相同。具体标签包括事件主体、事件原因、事件状态、事件时间和采取措施。一条日志在完成标注后,可以得到多条“SSI+Result+TXT”格式的数据。

2.5.2 SSI 和 DA-DeepSeek 提示语句的构造

本文数据增强模块与 UIE 抽取模块均采用 Prompt Learning 的思想。因此,合理构造 SSI 和 DA-DeepSeek 提示模板对本文方法至关重要。图 8 给出事件抽取过程中 Schema、SSI 及 DA-DeepSeek 提示模板的构造示例。

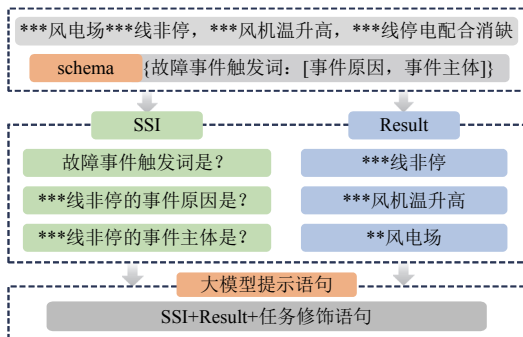


图 8 基于 SSI 的大模型提示语句构造示意图

Fig. 8 Schematic diagram of SSI-based prompt construction for large models

以图 8 中某风电场故障事件为例,构造的 SSI 是“***线非停的事件主体是?”, Result 是“***风电场”;对应的训练 DA-DeepSeek 提示语句为“***线非停的事件主体是***风电场,请帮我写出包含该主体的一条电力调度日志”,即“SSI+Result+任务修饰语句”;结合数据增强后的 TXT 文本,可以得到“SSI+TXT+Result”的数据格式,对 MSSI-UIE 模型的训练语料库进行扩充。

3 算例分析

3.1 实验设置与分析

为验证本文事件抽取方法的有效性,本文采用

我国某省级电网一年的调度日志进行分析,选取其中 736 条日志进行人工标注,转化为 12 248 条“SSI+Result+TXT”格式的数据,并以 6:2:2 构建训练集、测试集和验证集。实验所涉及的算法基于 Pytorch 深度学习框架搭建,训练环境基于 WSL2-Ubuntu 22.04.3 操作系统,Python 版本为 3.9, GPU 运行内存为 16 GB, MSSI-UIE 训练参数设置如表 1 所示。

表 1 训练参数设置

Table 1 Parameter setting of the training

参数	设置	说明
batch_size	16	批处理大小
learning_rate	10^{-5}	学习率
dropout_prob	0.1	注意力单元丢弃率
num_hidden_layers	12	transformer 编码器堆叠层数
epoch	100	训练轮数
hidden_act	gelu	隐藏层激活函数

选用精确率、召回率以及 F_1 值作为模型评价指标。3 个指标的计算方式为

$$P_{\text{recision}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \quad (9)$$

$$R_{\text{ecall}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \quad (10)$$

$$F_1 = \frac{2 \cdot P_{\text{recision}} \cdot R_{\text{ecall}}}{P_{\text{recision}} + R_{\text{ecall}}} \quad (11)$$

式中: N_{TP} 为真正例; N_{FP} 为假正例; N_{FN} 为假负例; P_{recision} 为计算模型预测出正确的样本占总预测样本的比例,反映模型预测实体的精确程度; R_{ecall} 为计算模型预测正确的样本占实际样本的比例,反映模型识别正确实体的能力。

精确率与召回率相互影响,召回能力越强,精确率通常伴随下降,反之亦然。因此,本文引入 F_1 值作为更全面的评价指标,用以评估 MSSI-UIE 模型的识别效果, F_1 值越高,表明模型对电力调度事件的抽取性能越好。

后续实验将在 BERT 和 ERNIE 两个基线预训练模型上开展性能评估,表 2 给出未经数据增强时这两个模型各自取得的 F_1 值。

由表 2 可以看出,“事件原因”这一事件论元在 BERT 与 ERNIE 两个基准预训练模型上的 F_1 值均低于其它事件论元,主要原因是并非所有的调度日志都包含“事件原因”这一事件论元,存在数据稀疏的问题,导致抽取效果较差。

表 2 基准预训练模型的 F_1 分数
Table 2 F_1 score of the baseline pre-trained model

事件论元	ERNIE_ F_1 /%	BERT_ F_1 /%
事件主体	88.88	83.33
事件原因	80.50	72.73
事件状态	91.66	87.54
事件时间	97.94	93.94
采取措施	90.90	85.71

3.2 数据增强结果分析

为了解决数据稀疏问题，生成符合电力调度日志文本特性的增强文本，本文基于 DeepSeek 和 Llama 2 种大模型进行数据增强实验。为了提高生成文本在格式和内容上的合理性，本文选择包含 Top-k、Top-p 和 Temperature 的联合采样策略，策略使用先后顺序是 Top-k、Top-p 和 Temperature。各采样策略的参数设置如表 3 所示。

表 3 DA-DeepSeek/DA-Llama 采样策略参数设置
Table 3 Parameter setting of
DA-DeepSeek/DA-Llama sampling strategy

采样策略	参数设置
Top-k 采样	50
Top-p 采样	0.95
Temperature 采样	0.3

综上可知，“事件原因”抽取的准确率不高，因此本节实例通过提示语句来生成包含“事件原因”的数据增强文本并分析其合理性。如表 4 所示，第 1 个实例中，DA-DeepSeek 生成的数据增强文本包含了“事件原因”即“55 号箱变高压熔断器熔断，造成弧光接地”，而 DA-Llama 所生成的数据增强文本“PCS-94D 保护过流 I 段动作”只是故障事件的一部分，并不是本文需要“事件原因”。第 2 个实例中，DA-Llama 所生成的数据增强文本包含的事件原因“线路#87-USB 灯故障，无法坚持运行”缺乏合理性，1 个 USB 灯故障，不会导致线路无法坚持运行，而且在原始的调度日志中，并没有 USB 故障这一事件原因，该结果由大模型幻觉导致。

表 4 DA-DeepSeek 与 DA-Llama 数据增强效果
Table 4 Data augmentation effects of
DA-DeepSeek and DA-Llama

DA-Llama 增强文本	DA-Deepseek 增强文本
*****线路跳闸 PCS-94D 保护过流 I 段动作，*****	*****，55 号箱变高压熔断器 熔断，造成弧光接地，*****
*****，线路#87-USB 灯故障， 无法坚持运行，*****	*****山火导致 B 相 跌落保险熔断，*****

DA-DeepSeek 生成的数据优于 DA-Llama，主

要归因于两方面：其一，DeepSeek 在基础模型上经过大规模强化学习，具备了思维链(CoT)推理能力，使其输出结果更具可解释性和合理性；其二，DeepSeek R1 针对中文文本进行了专门训练，相较于英文为主的 Llama，其对中文语义的理解更为精准，生成内容也更符合中文表达习惯。

实验结果表明，DA-DeepSeek 可以得到格式和内容均较为合理的增强文本，本文利用 DA-DeepSeek 共生成 7 449 条增强文本，以扩充 MSSI-UIE 模型的训练语料库。

为全面验证 MSSI-UIE 方法的有效性，本文针对 MSSI-UIE 方法的不同模块分别开展对比实验。

3.3 事件抽取性能分析

3.3.1 数据增强效果分析

首先，对比 ERNIE 和 BERT 2 种预训练模型作为 UIE 的编码器，在数据增强前后的事件抽取效果，以验证数据增强的有效性以及不同预训练模型对 UIE 抽取性能的影响。

实验结果如表 5 所示。通过数据增强，ERNIE 预训练模型的 F_1 值提升了 2.16%，BERT 预训练模型提升了 6.68%。未经数据增强时，ERNIE 预训练模型在这类任务中的性能表现远超 BERT。因为 ERNIE 模型由百度在通用数据集上微调得到，能凭借其预训练积累的知识，在未经数据增强的情形下展现出良好的抽取性能；而 BERT 预训练模型未经过此类任务的再训练，其抽取性能较差。进行数据增强后，训练语料库数据量成倍增长，模型得以充分训练，对通用知识的依赖程度降低。因此，BERT 和 ERNIE 这两个模型在抽取性能上的差距急剧缩小。这表明数据增强能有效提升模型性能，且数据增强后 BERT 预训练模型具有更好的抽取效果。

表 5 不同编码器与数据增强前后指标对比

Table 5 Metrics comparison of
different encoders before and after data augmentation

模型	精确率/%	召回率/%	F_1 /%
ERNIE 数据增强前	90.06	89.47	89.76
ERNIE 数据增强后	89.51	94.46	91.92
BERT 数据增强前	86.98	83.53	85.56
BERT 数据增强后	90.49	94.0	92.24

为更直观地展示数据增强带来的事件抽取性能的提升，对调度日志实例 1 “*****，110 kV***线跳闸，*****大型飞禽穿越***杆 B 相线路时，导致 B 相接地短路*****”与实例 2 “***电场 110 kV***线跳闸，*****，C 相故障”在数据增

强后的事件抽取情况进行对比，如表 6 所示。

表 6 数据增强后模型抽取效果验证

Table 6 Validation of model extraction effectiveness after data augmentation		
模型	实例 1 抽取结果	实例 2 抽取结果
BERT	(110 kV***线跳闸, 事件原因, B 相接地短路)	(110 kV***线跳闸, 事件原因, C 相故障)
ERNIE	(110 kV***线跳闸, 事件原因, B 相接地短路)	—

数据增强前，两个实例的事件原因均未能有效抽取。数据增强后(如表 6 所示)，BERT 预训练模型对两个实例均能准确抽取事件原因；而 ERNIE 预训练模型仅能抽取实例 1，对实例 2 仍无法有效抽取。这一结果验证了本文 DA-DeepSeek 数据增强方法的有效性，同时也表明，训练语料扩充对 BERT 模型的提升效果更为显著。

为了更好地分析数据增强在模型训练过程中的作用，本节对比了两种预训练模型在数据增强前后的 F_1 值变化曲线，如图 9 所示。使用 ERNIE 预训练模型作为编码器时，可以在训练初期就取得较高的 F_1 值；使用 BERT 预训练模型作为编码器时，在前 10 个训练轮次中无法抽取出正确信息，在 10~19 轮 F_1 值快速提升至 0.75，然后缓慢提升，在训练 100 轮次时， F_1 值接近 ERNIE 预训练模型。数据增强后，两种预训练模型的 F_1 值均有所提高，且模型的训练收敛速度也明显加快。训练语料库扩充后，BERT 预训练模型效果提升更加显著，且 F_1 值优于 ERNIE 预训练模型，因此本文选用 BERT 预训练模型作为 UIE 的编码器。

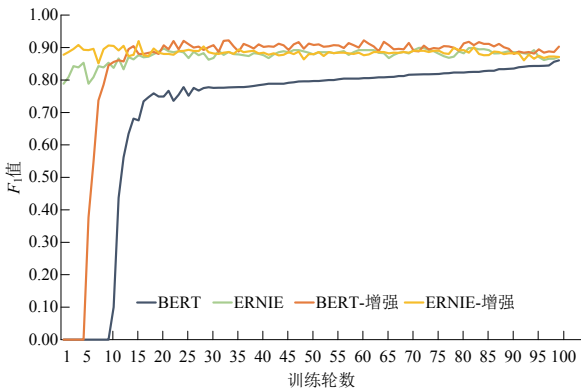


图 9 对比模型的训练监控指标对比

Fig. 9 Comparison of training monitoring metrics for contrastive models

3.3.2 电力实词 BERT 预训练实验

为进一步提高 BERT 预训练模型的抽取能力，本文基于 ERNIE 实词掩码思想对 BERT 进行动态实

词掩码训练^[20]，即将电力调度日志中的设备名称、故障类型、检修操作等高频词汇对 BERT 进行实体与短句掩码预训练。完成预训练后，再次对 BERT 进行 UIE 事件抽取任务训练，表 7 展示了 BERT 模型在经过实词掩码预训练和数据增强后的性能。

表 7 电力实词预训练前后各指标对比

Table 7 Comparison of metrics before and after pretraining on power-domain terms			
模型	精确率/%	召回率/%	F_1 /%
BERT	90.49	94.07	92.24
BERT +实词预训练	91.57	94.13	93.00

如表 7 所示，BERT 模型经过电力文本的实词掩码预训练后，精确率提升 1.08%、召回率提升 0.36%， F_1 值提升 0.76%。

3.3.3 MSSI 联合抽取实验

为验证 MSSI 联合抽取的有效性，本节对比了事件论元抽取、事件三元组抽取及 MSSI 联合抽取 3 种方式在通用型论元和事件原因论元上的 F_1 值。

由表 8 可知，本文提出的 MSSI 联合抽取方法相较于单独的通用型事件论元抽取和事件三元组抽取，能够有效提升事件抽取性能。其中，整体 F_1 值提升 1.04%，事件原因论元抽取的 F_1 值提升尤为显著，达 6.51%。

表 8 MSSI 联合抽取对 F_1 值提升

Table 8 Improvement in F_1 score through joint MSSI extraction		
抽取方式	整体 F_1 /%	事件原因 F_1 /%
通用型事件论元抽取	93.15	84.85
事件三元组抽取	92.64	76.13
MSSI 联合抽取	94.19	91.36

3.3.4 主实验

为进一步验证 MSSI-UIE 的有效性，将其与 BiLSTM-CRF^[10]和 BERT-BiLSTM-CRF^[11]2 种传统抽取方法进行了对比试验，结果如表 9 所示，MSSI-UIE 方法取得了最优的抽取性能。

表 9 MSSI-UIE 与其余模型的对比

Table 9 Comparison of MSSI-UIE with other models			
方法	精确率/%	召回率/%	F_1 /%
BiLSTM-CRF	80.58	82.45	81.50
BERT- BiLSTM-CRF	84.33	87.73	85.99
MSSI-UIE	93.38	95.03	94.19

具体来说，MSSI-UIE 对比 BERT- BiLSTM-CRF 模型在精确度上提升 9.05%，在召回率上提升 7.30%， F_1 值上提升 8.20%。比 BiLSTM-CRF 模型

在精确度上提升 12.58%，在召回率上提升 12.58%， F_1 值上提升 12.69%。MSSI-UIE 在事件抽取任务中表现优于其他方法，主要原因包括：

1) 数据增强：采用 DA-DeepSeek 生成高质量的电力调度增强文本，提升了模型的泛化能力；

2) 领域自适应预训练：基于电力调度文本对 BERT 进行实词掩码预训练，增强了模型对电力专业术语的理解；

3) 联合抽取优化：MSSI 方法通过多任务协同学习，提高了事件抽取的准确性。

4 结论

本文提出了一种基于 MSSI-UIE 的电力调度事件抽取方法，显著提升了在小样本场景下电力调度事件抽取的可靠性。本文以数据增强后的 19 597 条电力调度文本，从多个角度设计并进行了大量实验，验证了 MSSI-UIE 方法相比于其他方法性能的优越性，探究了模型中各个模块的作用和有效性，以及在电力稀疏文本以及小样本场景下的事件抽取提升效果，相比 BERT-BiLSTM-CRF 等方法有至少 8.20% 的提升，为准确抽取电力调度事件文本中的经验知识，构建电力调度知识图谱奠定了基础。基于 MSSI-UIE 的事件抽取方法主要结论如下：

1) 结合电力调度文本特性，提出基于提示学习以及事件抽取任务转换的 UIE 事件抽取方法，并对 BERT 编码器进行电力实词掩码预训练，实现电力调度事件的抽取。

2) 提出基于大模型的数据增强方法，将 SSI 与大模型提示语句结合，形成完整的数据增强工作流程，提升模型的事件抽取能力。

3) 提出多结构化提示器 MSSI 联合抽取方法，该方法通过多任务协同学习，充分利用调度文本中的语义信息，提高对稀疏样本事件抽取的可靠性。

本文所提出的事件抽取方法基于某省级电网调度日志中的事件文本进行实验验证。由于不同地区电网调度日志所涉及的设备名称和调度员表述习惯存在差异，本方法的跨区域适用能力仍有待进一步研究。

参考文献

- [1] 乔骥, 王新迎, 闵睿. 面向电网调度故障处理的知识图谱框架与关键技术初探[J]. 中国电机工程学报, 2020, 40(18): 5837-5848.
- QIAO Ji, WANG Xinying, MIN Rui. Framework and key

technologies of knowledge-graph-based fault handling system in power grid[J]. Proceedings of the CSEE, 2020, 40(18): 5837-5848(in Chinese).

- [2] 张智刚, 康重庆. 碳中和目标下构建新型电力系统的挑战与展望[J]. 中国电机工程学报, 2022, 42(8): 2806-2818.

ZHANG Zhigang, KANG Chongqing. Challenges and prospects for constructing the new-type power system towards a carbon neutrality future[J]. Proceedings of the CSEE, 2022, 42(8): 2806-2818(in Chinese).

- [3] 张禹方, 袁之康, 高硕杰, 等. 基于跨模态数据的变压器套管故障知识图谱构建与应用[J]. 中国电机工程学报, 2025, 45(22): 9064-9074.

ZHANG Yufang, YUAN Zhikang, GAO Shuojie, et al. Construction and application of knowledge graph of transformer bushing faults based on cross-modal data[J]. Proceedings of the CSEE, 2025, 45(22): 9064-9074(in Chinese).

- [4] 齐冬莲, 闫玮丹, 闫云凤, 等. 面向电力调度的事件知识图谱研究现状及发展[J]. 电子与信息学报, 2024, 46(9): 3456-3466.

QI Donglian, YAN Weidan, YAN Yunfeng, et al. A review of research methods on event knowledge graph for power dispatching[J]. Journal of Electronics & Information Technology, 2024, 46(9): 3456-3466(in Chinese).

- [5] 李鹏, 余涛, 李立涅, 等. 电力人工智能的演变与展望——从专业智能走向通用智能[J]. 电力系统自动化, 2024, 48(16): 1-17.

LI Peng, YU Tao, LI Licheng, et al. Retrospect and prospect of artificial intelligence for electric power system — from domain intelligence to general intelligence[J]. Automation of Electric Power Systems, 2024, 48(16): 1-17(in Chinese).

- [6] 郭榕, 杨群, 刘绍翰, 等. 电网故障处置知识图谱构建研究与应用[J]. 电网技术, 2021, 45(6): 2092-2100.

GUO Rong, YANG Qun, LIU Shaohan, et al. Construction and application of power grid fault handling knowledge graph[J]. Power System Technology, 2021, 45(6): 2092-2100(in Chinese).

- [7] 田生伟, 周兴发, 禹龙, 等. 基于双向 LSTM 的维吾尔语事件因果关系抽取[J]. 电子与信息学报, 2018, 40(1): 200-208.

TIAN Shengwei, ZHOU Xingfa, YU Long, et al. Causal relation extraction of Uyghur events based on bidirectional long short-term memory model[J]. Journal of Electronics & Information Technology, 2018, 40(1): 200-208(in Chinese).

- [8] LI Qian, LI Jianxin, SHENG Jiawei, et al. A survey on deep learning event extraction: approaches and applications[J]. IEEE Transactions on Neural Networks

- and Learning Systems, 2024, 35(5): 6301-6321.
- [9] CHEN Yubo, LIU Shulin, HE Shizhu, et al. Event extraction via bidirectional long short-term memory tensor neural networks[C]//15th China National Conference on Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Yantai: Springer, 2016: 109-203.
- [10] 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).
- [11] 蒋晨, 王渊, 胡俊华, 等. 基于深度学习的电力实体信息识别方法[J]. 电网技术, 2021, 45(6): 2141-2149.
JIANG Chen, WANG Yuan, HU Junhua, et al. Power entity information recognition based on deep learning[J]. Power System Technology, 2021, 45(6): 2141-2149(in Chinese).
- [12] 万齐智, 万常选, 胡蓉, 等. 面向研究问题的深度学习事件抽取综述[J]. 自动化学报, 2024, 50(11): 2079-2101.
WAN Qizhi, WAN Changxuan, HU Rong, et al. Event extraction based on deep learning: a survey of research issue[J]. Acta Automatica Sinica, 2024, 50(11): 2079-2101(in Chinese).
- [13] 谢庆, 蔡扬, 谢军, 等. 基于 ALBERT 的电力变压器运维知识图谱构建方法与应用研究[J]. 电工技术学报, 2023, 38(1): 95-106.
XIE Qing, CAI Yang, XIE Jun, et al. Research on construction method and application of knowledge graph for power transformer operation and maintenance based on ALBERT[J]. Transactions of China Electrotechnical Society, 2023, 38(1): 95-106(in Chinese).
- [14] 戴志辉, 张富泽, 张近月, 等. 基于 MacBERT-BiLSTM-CRF 模型的继电保护装置缺陷知识图谱构建方法[J]. 电力系统保护与控制, 2024, 52(20): 131-143.
DAI Zhihui, ZHANG Fuze, ZHANG Jinyue, et al. Construction method of a defect knowledge map of a relay protection device based on a MacBERT-BiLSTM-CRF model[J]. Power System Protection and Control, 2024, 52(20): 131-143(in Chinese).
- [15] DU Xinya, CARDIE C. Event extraction by answering (almost) natural questions[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online: Association for Computational Linguistics, 2020: 671-683.
- [16] LI Fayuan, PENG Weihua, CHEN Yuguang, et al. Event extraction as multi-turn question answering[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, 2020: 829-838.
- [17] LU Yaojie, LIU Qing, DAI Dai, et al. Unified structure generation for universal information extraction[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: Association for Computational Linguistics, 2022: 5755-5772.
- [18] BARADARAN R, GHIASI R, AMIRKHANI H. A survey on machine reading comprehension systems[J]. Natural Language Engineering, 2022, 28(6): 683-732.
- [19] DEVLIN J, CHANG Mingwei, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.
- [20] LIU Yinhan, OTT M, GOYAL N, et al. ROBERTA: a robustly optimized BERT pretraining approach[C]//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2020: 1-15.



谭琦

在线出版日期: 2025-09-09。

收稿日期: 2025-05-12。

作者简介:

谭琦(1985), 男, 博士, 讲师, 硕士生导师, 研究方向为人工智能技术在电力系统优化调度中的应用, tanqi@hfut.edu.cn;

朱道坤(2000), 男, 硕士研究生, 研究方向为电力领域知识图谱的构建与应用, 1203798736@qq.com;

唐昊(1972), 男, 博士, 教授, 博士生导师, 研究方向为离散事件动态系统、强化学习及深度(强化)学习等智能优化与控制方法, htang@hfut.edu.cn。

(责任编辑 刘雪莹)