第 50 卷 第 11 期：4912-4924 高电压技术 Vol.50, No.11: 4912-4924
2024 年 11 月 30 日 High Voltage Engineering November 30, 2024
DOI: 10.13336/j.1003-6520.hve.20230772

面向电力知识图谱构建的重叠实体关系联合抽取方法

束嘉伟，杨　挺，耿毅男，于　洁
（智能电网教育部重点实验室(天津大学)，天津 300072）

摘　要：作为构建电力知识图谱的关键步骤，知识抽取可以从海量非结构化电力文本中准确抽取出实体和关系。但是，传统流水线式方法存在识别的错误信息后向传递、实体识别和关系抽取任务割裂以及易产生冗余信息的问题，进而导致抽取准确率低、抽取信息不全面，最终影响知识图谱的准确构建。针对上述问题，提出面向电力知识图谱构建的重叠实体关系联合抽取方法，通过改进的序列标注方案进行联合抽取，构建了电力领域专属预训练 PowerRoberta 模型，并增加对抗训练，提高了模型抽取电力知识的准确度和对陌生信息的预测能力。最后，以实际变电站巡检数据为例进行了实验分析与配电知识图谱可视化构建，结果表明所提出的联合抽取方法提升了知识抽取的准确率，准确率达到 91.67%，可有效支撑配电网智能信息检索、辅助决策高级应用。
关键词：自然语言处理；电力知识图谱；知识抽取；实体关系联合抽取；序列标注；关系重叠

Joint Extraction Method for Overlapping Entity Relationships in the Construction of Electric Power Knowledge Graph

SHU Jiawei, YANG Ting, GENG Yinan, YU Jie
(Smart Grid Key Laboratory of the Ministry of Education (Tianjin University), Tianjin 300072, China)

Abstract：As a key step in building a power knowledge graph, knowledge extraction can accurately extract entities and relationships from massive unstructured power texts. However, the traditional pipeline method has the problems of backward transmission of error information, separation of entity recognition, and relationship extraction tasks, and is easy to generate redundant information, which results in low extraction accuracy, incomplete extraction of information, and ultimately impairs the accurate construction of the knowledge graph. To solve the above problems, this paper proposes a joint extraction method of overlapping entity relationships for the construction of the power knowledge graph. Through the improved sequence labeling scheme, the joint extraction is carried out, the exclusive pre-training model (the PowerRobertsa model) in the power field is constructed, and the confrontation training is increased, which improves the accuracy of the model extraction of power knowledge and the ability to predict unfamiliar information. Finally, by taking the actual substation patrol data as an example, the experimental analysis and the visual construction of the distribution Knowledge graph are carried out. The results show that the joint extraction method proposed in this paper can be adopted to improve the accuracy of knowledge extraction, which reaches 91.67%, and can effectively support the advanced application of distribution network intelligent information retrieval and decision-making assistance.
Key words：natural language processing; power knowledge graph; knowledge extraction; joint extraction of entity relations; sequence labeling; relationship overlap

0　引言

随着大数据时代的到来，人工智能技术促进了智能电网的发展和应用，电网每天产生的数据量也在急剧膨胀，其中配电网运行及巡检数据包括设备运行状态、设备故障报告和缺陷分析、检修记录、管理规定等非结构化的文本类数据[1]，存在数据利用率低、数据间的相关性缺失等严重问题。伴随着人工智能技术的不断深入发展，知识图谱作为一种数据挖掘技术，能够把海量数据中蕴含的复杂知识可视化地呈现，为知识规律的研究提供了有力参考，受到了各国学者的重视。因此，搭建成熟完整的电力知识图谱，既是有效描述电力领域技术体系、助力电力行业相关技术人员对电力知识进行系统化和结构化管理的关键，也是电力领域向智能化管理运营转型升级的迫切需求。

基金资助项目：国家重点研发计划(2022YFB2403800)；国家自然科学基金(U2066213)；天津市自然科学基金重点项目(21JCZDJC00640)。
Project supported by National Key R&D Program of China (2022YFB2403800), National Natural Science Foundation of China (U2066213), Natural Science Foundation Key Project of Tianjin (21JCZDJC00640).

知识图谱的构建过程包括数据获取、知识抽取、知识融合、知识加工。知识抽取是构建知识图谱最核心的环节[2-4]，但由于自然语言语义的复杂性和模糊性，针对电力文本的分析和抽取不仅要考虑文字本身的含义，还需结合电力行业背景、专业术语以及特定表达等实际因素进行理解抽取。因此搭建有效的电力知识抽取模型，是提升抽取准确率，挖掘非结构化文本深层隐含信息的关键，更是构建完整准确的电力知识图谱、进而辅助电力技术人员高效运维决策的重要步骤，具有重要意义。

知识抽取通常依赖 2 个下游任务，即命名实体识别[5-7]和关系抽取[8-10]。目前多数研究学者采取传统的流水线方法，即将命名实体识别和关系抽取划分为 2 个独立的子任务逐步处理，首先进行命名实体识别，而后将实体模型提取出的实体进行关系配对，最后实现关系分类。流水线方法虽然在灵活性和独立性方面有优势，但存在以下 3 方面问题：(1)错误信息传播。命名实体识别的效果直接影响关系抽取的效果，上游任务产生的错误信息会传播到关系抽取任务中，存在误差积累和传播识别产生的错误信息的问题，进而影响关系抽取结果。(2) 2 个子任务之间缺乏互动。流水线方法忽略了实体识别和关系抽取这 2 个子任务之间的相关性，割裂了其内在联系，导致实体之间隐含的关系难以被发掘，影响最终抽取效果。(3)产生冗余信息。流水线式是通过命名实体识别提取出实体信息，而并非所有实体之间都存在一定的关系，部分不存在关系的独立实体传递到关系抽取任务就成为冗余信息，影响分类效果，增大了提取错误率。因此，实体与关系的联合抽取越来越受到研究人员的重视。

联合抽取方法通过一个模型同时进行实体识别和关系抽取[11]，增加了 2 个任务的耦合性，具有更好的抽取表现。目前一部分学者采用基于共享参数[12-13]的联合抽取方法，该方法通过共享参数的方法实现信息抽取，能够解决流水线式方法中错误传播以及信息冗余的问题。然而，以上方法并非真正意义上的联合，仍存在子任务间交互性不强的问题，难以精准地抽取实体关系、有效挖掘隐含信息，导致构建的知识图谱信息不完整。

为了解决以上问题，部分学者做了初步探究，发现基于序列标注和 Sequence-to-Sequence 序列到序列联合解码的联合抽取模型可以很好地实现不同子任务间的交互。其中，Sequence-to-Sequence 方法存在复杂句子抽取难度大和信息冗余的问题，序列标注方法通过设计统一的解码器实现了实体关系信息充分交互[14-15]，但常见的序列标注方法无法解决重叠关系、实体嵌套的问题。

现有研究无法完全解决知识图谱构建过程存在的错误信息传播、产生冗余信息和重叠实体关系等问题，且针对配电网的知识图谱构建及应用研究同样尚未深入开展。因此本文面向配电网运行及巡检文本数据，提出一种基于改进的序列标注联合抽取方法，搭建 PowerRoberta-PGD-BiLSTM-CRF (PPBC)深度学习模型进行电力图谱知识抽取。

同时，本文参考省级电网公司巡检文本记录和电力设备缺陷记录，机器学习并标注了 200 篇电力行业国家标准、企业标准和规章制度文档，如《电工术语 发电、输电及配电 电力系统规划和管理》、《变电站标准化管理条例》、《变电站运行管理标准》、电力设备运行规程及管理规定、电力设备技术标准等，对原始电力语料进行更准确的手工标注和模型训练，最后将抽取的知识三元组利用 Neo4j 图数据库进行存储和可视化表示，构建一个用于配电网智能检索和辅助决策的知识图谱。

因此，本文基于配电网运行及巡检文本特征提出一种针对电力实体关系联合抽取的序列标注策略：“实体位置+实体类型+关系类别+实体角色”(BIESO+Entity+Relation+Role)。将构建电力知识图谱的关键步骤——命名实体识别和关系抽取任务转化为序列标注问题，将输入的电力文本序列针对实体和关系在同一标签下进行标注，直接对三元组建模进行实体和关系联合抽取，解决了实体冗余和重叠关系抽取问题，避免了错误传播，显著提升电力领域实体关系提取和标注的准确率和效率，实现了更全面更准确的电力知识图谱构建。本文的主要工作如下：

1）针对电力文本中实体分布密集且关系重叠的难点，设计了实体关系联合抽取模型。模型将联合抽取任务转化为序列标注问题，进而提出一种基于“BIESO+Entity+Relation+Role”的标注策略。新标注策略突破了传统序列标注中单个实体只能隶属于一种关系而无法抽取重叠关系的局限性，并且实现了实体标注与关系抽取的同步完成，从而有效减少了实体冗余度和错误信息后向传递，提升了电力知识图谱构建过程中的标注效率和抽取准确度。

2）通过对实际配电网运行/巡检文本数据分析，

及对电力领域国标/团标和电力企业规章制度文档机器学习，本文构建了电力领域 PowerRoberta 预训练模型，并加入到实体关系联合抽取模型中，提升模型识别实体边界的准确性。进而使用 PowerRoberta-PGD-BiLSTM-CRF 组合模式实现电力文本全局上下文特征的准确提取，将抽取得到的知识三元组存储在 Neo4j 图数据库中，搭建出电力知识图谱。

3）本文在抽取过程中引入投影梯度下降(projected gradient descent，PGD)对抗训练，通过在 PowerRoberta 模块训练得到的参数矩阵中注入小干扰，增强对电力实体关系抽取模型的抗干扰能力和对未知信息的预测能力，有效提升了电力实体关系抽取的准确率和鲁棒性。

1 联合抽取标注策略

为了解决基于共享参数的联合抽取模型信息交互不充分的问题，相关研究往往通过序列标注方法，设计统一的解码器，编码后的句子序列直接被解码得到序列标注信息，从而获得关系三元组，但无法解决关系重叠问题。实际中的电力文本往往存在复杂的实体关系重叠情况，如一个实体与其他实体之间存在多种关系，即：“欧式箱变”和生产厂商存在生产关系，同时和变电站存在运行关系。针对此问题，本文将构建电力知识图谱的联合抽取任务转化为序列标注问题，同时提出了考虑复杂实体关系重叠情况的序列标注策略，通过添加实体关系类别并丰富实体角色，解决了传统序列标注中重叠关系抽取问题，实现对实体和关系类型的精准预测，有效提升了抽取的准确率，更有利于搭建准确专业的电力知识图谱。

本文定义标签由 4 部分组成：“实体位置-实体类型-关系类别-实体角色”。首先采用“BIESO”标签来表示实体中词的位置边界信息，具体含义如表 1 所示。

同时，本文在实体位置边界标签的基础上添加了实体类别和关系类别标签，以体现实体间的关系。另外新定义一种标签类型“Overlapping”(简记为“OL”)，用来标记重叠关系，根据对应的重叠实体和关系进行匹配。因此，重叠情况的标签可由“实体位置信息-实体类型-OL-关系角色”构成，解决序列标注中出现的重叠关系的抽取问题。本文标注的实体类型和关系类别标签如表 2 所示。

例如，选取配电设备资产数据库某一语料：“海光寺变电站的变压器是欧式箱变，由天津电力设备公司生产”，其包含 2 组关系：①运行/工作关系(OT)：“海光寺变电站”包含“欧式箱变”；②生产关系(PD)：“欧式箱变”的生产厂家为“天津电力设备公司”。标注策略示例如图 1 所示。

在本例中，实体“欧式箱变”同属于上述 2 关系，即存在关系重叠，因此采用本文引入的“OL”重叠关系标签代替表示“欧式箱变”(B/I/E-EQU-OL-HT)，其中标签“HT”清晰地表明了该实体为重叠实体，既是前一运行关系“OT”的头实体，又是后一生产关系“PD”的尾实体。

其中“H”表示在三元组中处于头实体(Head)，“T”则表示处于尾实体(Tail)，形成实体的有向性表示。而针对电力文本中长程语句造成的重叠实体情况，可采用组合表示，如“TH”表示该实体是前一个三元组的尾实体，同时是后一个三元组的头实体，如图 2(a)所示。此外还有如图 2(b)所示的“HT”情况，含义与“TH”表示相反。

在电力领域中，并非所有的词都是实体的组成部分，也并非所有实体都参与关系，且对于变电站的名称和位置等属性，流水线式抽取时可能进行多次抽取，同一实体频繁出现，存在重复信息；同时也会将不存在的关系属性及实体抽取出来。而序列标注针对不同实体类型和属性定义过滤器和规则，过滤掉不必要的标记，避免重复信息的出现。因此，

表 1 实体位置边界标签含义

Table 1 Meaning of entity location boundary labels

位置边界标签	标签含义
B	Begin，表示实体的开始
I	Inside，表示实体的中间位置
E	End，表示实体的结束
S	Single，表示单一实体
O	Other，不在抽取范围之内的其他类实体

表 2 实体类型和关系类别标签

Table 2 Entity types and relationship category labels

标签类型	标签内容	标签含义
实体类型	LOC	地点
	EQU	设备
	FAU	故障
	MF	生产厂商
	VAL	数值/等级
关系类别	OT	运行/隶属关系
	PD	生产关系
	SL	解决/处理关系
	OL	重叠关系

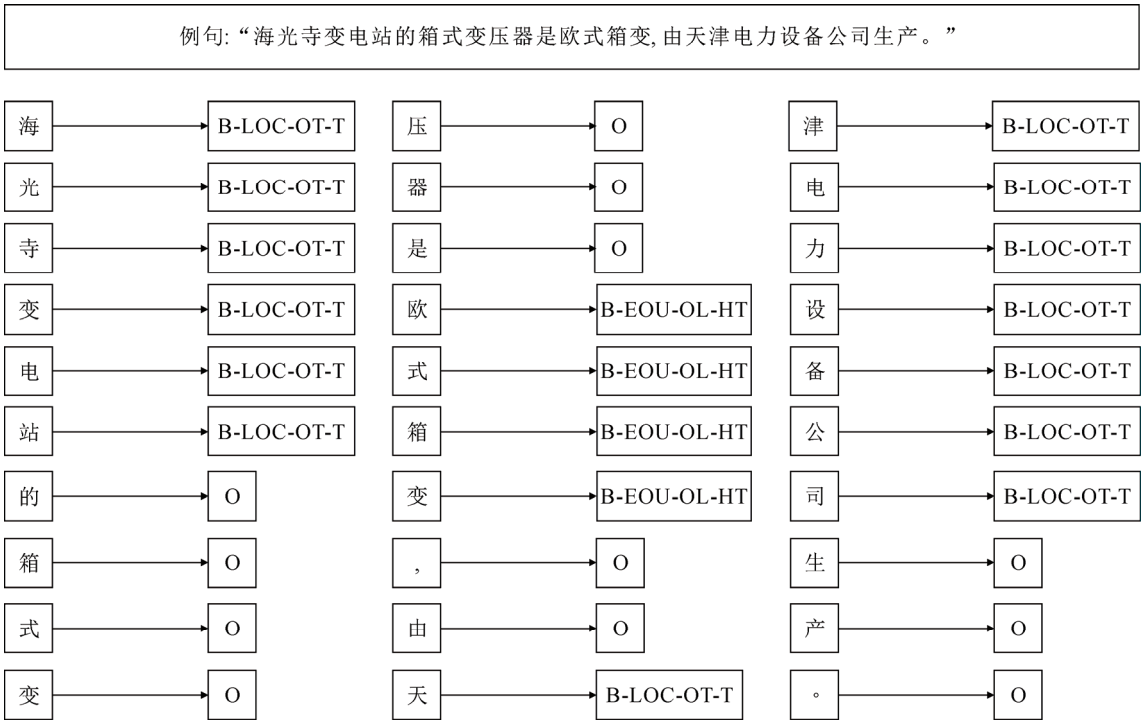


图 1　标注策略示例图

Fig.1　Example diagram of annotation strategy

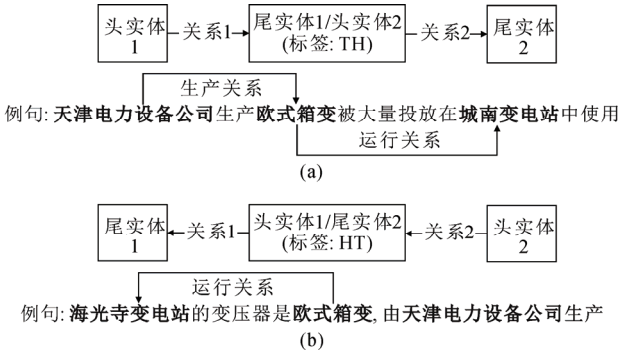


图 2　实体关系抽象图

Fig.2　Abstract diagram of entity relationships

为了降低标签的复杂度和冗余度，本文设计出 3 种实际类型的标签格式：(1)非实体标签“O”；(2)只有实体类型不参与关系，如单一实体+地点的“S-LOC”格式；(3)实体和关系都存在，如实体-生产厂家-生产关系-头实体“B-MF-PD-H”。

此外，对于错误信息后向传递问题，例如抽取某一语料：“天津变压器厂生产的 10 kV 变压器变 380 V 型号为 S9-630/10”。可能抽取出实体“变压器变 380 V”，其并非具体的变压器名称，而是二次侧额定电压等级。但流水线式方法则会将该错误命名实体识别传递给下一级关系抽取任务，导致错误名称与其他实体和关系进行匹配。而序列标注能够同时抽取如变压器名称、生产厂商、型号、额定电压等级等多个信息，因此避免了错误信息后向传递弊端。

2　基于 PowerRoberta-PGD-BiLSTM-CRF 的电力实体关系联合抽取模型

2.1　PowerRoberta 层

注意到中文表述中词之间并没有像英语的分隔符，通用领域的传统 BERT 模型在使用中文电力语料进行预训练时，只能掩盖字而不能掩盖词，因此无法获得词级别的语义表示，即无法在预训练过程中获取电力领域词的信息。

为了在模型训练中更全面地抽取电力文本词的信息，本文采用 Roberta 预训练语言模型。Roberta 使用了全词掩码策略，能够完整获取中文电力文本中词的信息，提高对电力文本粗粒度语义的建模能力，从而显著提升模型性能。同时 Roberta 使用动态掩码策略，每次向模型输入一个序列时都生成新的掩码模式，动态地改变应用于训练数据的遮蔽模式。与传统的静态词向量训练相比，Roberta 模型训练输出的动态词向量包含了词的含义和上下文词之间的特征，不断适应时刻变化的掩码策略，能够更灵活地学习电力语料特征，更适用于中文电力领域

的实体关系抽取任务。

但是通用领域的词分布与电力领域存在较大差异，电力文本中通常存在分布更密集的专业词汇，因此难以利用通用语言模型进行学习理解。针对上述问题，本文使用主变压器、断路器与隔离开关 3 类电力一次设备的相关巡检文本记录和电力设备缺陷记录、试验或故障分析报告以及电力行业国家标准、企业标准和规章制度文档(如：GB/T 50297—2018 电力工程基本术语标准、GB/T 2900.1—2008 电工术语 基本术语、GB/T 2900.58—2008 电工术语 发电、输电及配电电力系统规划和管理、GB/T 32893—2016 10 kV 及以上电力用户变电站运行管理规范、QCVN QTD-5—2009/BCT 电力技术技术标准第 5 部分：电力系统设备校准、《变电站标准化管理条例》、《电力设备运行规程及管理规定》等)对 Roberta 模块进行了预训练，构建出专属电力领域的 PowerRoberta 模型，模型架构及输入表述如附录 A 图 A1 所示。

利用 PowerRoberta 预训练模型中的 Transformer 结构，对输入电力文本序列进行编码，使用自注意力机制训练得到对应的向量。对于输入的电力文本序列 $X=\{x_1, x_2, x_3, \cdots, x_n\}$，$n$ 为词的个数，将每个词采用独热向量(one-hot)表示，设维度为 k，则输入序列对应的嵌入矩阵为 $\boldsymbol{A}_n$。多头注意力层的输入为矩阵 $\boldsymbol{A}_n$，Roberta 自注意力层输入 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 的计算式为：

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{A}_n\boldsymbol{W}^{\boldsymbol{Q}}\\ \boldsymbol{K}=\boldsymbol{A}_n\boldsymbol{W}^{\boldsymbol{K}}\\ \boldsymbol{V}=\boldsymbol{A}_n\boldsymbol{W}^{\boldsymbol{V}}\end{cases} \tag{1}$$

式中：$\boldsymbol{W}^{\boldsymbol{Q}}$、$\boldsymbol{W}^{\boldsymbol{K}}$、$\boldsymbol{W}^{\boldsymbol{V}}$ 为权重矩阵。

自注意力层输出式为：

$$\boldsymbol{A}_{\text{attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=S_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

式中：d_k 为 $\boldsymbol{Q}$、$\boldsymbol{K}$ 的列数，即向量的维度；$S_{\text{softmax}}(\bullet)$ 为归一化函数。

Transformer 使用多头注意力机制 Multi-head Self-attention，输出矩阵 $\boldsymbol{M}_{\text{multiHead}}$ 的计算式为：

$$\boldsymbol{M}_{\text{multiHead}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=C_{\text{concat}}\left(\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3,\cdots,\boldsymbol{H}_\mu\right)\boldsymbol{W}^{\boldsymbol{O}} \tag{3}$$

$$\boldsymbol{H}_\beta=\boldsymbol{A}_{\text{attention}}\left(\boldsymbol{Q}\boldsymbol{W}_\beta^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_\beta^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_\beta^{\boldsymbol{V}}\right) \tag{4}$$

式中：μ 表示头的数量；$\boldsymbol{W}_\beta^{\boldsymbol{Q}}$、$\boldsymbol{W}_\beta^{\boldsymbol{K}}$、$\boldsymbol{W}_\beta^{\boldsymbol{V}}$ 分别表示第 β 个头的 $\boldsymbol{W}^{\boldsymbol{Q}}$、$\boldsymbol{W}^{\boldsymbol{K}}$、$\boldsymbol{W}^{\boldsymbol{V}}$ 权重矩阵；$\boldsymbol{W}^{\boldsymbol{O}}$ 表示附加的权重矩阵；$C_{\text{concat}}(\bullet)$ 表示拼接函数；$\boldsymbol{H}_\beta$ 表示每个头的输出矩阵。

PowerRoberta 模型将输入的电力文本序列转化为表示词向量、句向量以及位置量，进而输入到模型中进行提取，利用自注意力机制和全连接层对输入的电力文本进行建模，从大量文本信息中选择出和当前任务目标更契合的有效信息。为了进一步对电力文本信息进行特征学习，将巡检文本以句子为单位输入进已预训练的 PowerRoberta 模块以得到高质量的词向量，再利用 BiLSTM 层学习电力文本的上下文特征。

2.2 BiLSTM 层

通过大量的实践应用发现，深度学习长短期记忆网络(long short-term memory，LSTM)可以实现文本上下文信息的有效存储和更新，保持数据之间的依赖关系。但是 LSTM 无法从尾到头编码信息，双向长短期记忆网络(bidirectional long short-term memory，BiLSTM)模型则可以获得双向语义依赖并获得更全面的文本信息。BiLSTM 层中 LSTM 记忆单元用于计算当前的隐藏向量 $\boldsymbol{h}_t$，在前一时刻隐藏向量 $\boldsymbol{h}_{t-1}$ 基础上，前一时刻单元向量 $\boldsymbol{c}_{t-1}$ 和当前的输入词嵌入 $\boldsymbol{w}_t$。BiLSTM 模型是通过输入门、遗忘门和输出门来控制信息的流动和计算处理。PowerRoberta 层将电力文本转换为嵌入向量，因此一个词的序列可以表示为 $\boldsymbol{X}^*=[\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \cdots, \boldsymbol{x}_\eta^*]$，其中 $\boldsymbol{x}_t^*\in\mathbf{R}^d$ 是对应于经过 PowerRoberta 层处理后的句子中第 t 个词的 d 维词向量，$\mathbf{R}$ 为实数范围，η 表示给定句子的长度。BiLSTM 的具体训练式为：

$$\boldsymbol{i}_t=\sigma\left(\boldsymbol{w}_{x_{\text{i}}}\boldsymbol{x}_t^*+\boldsymbol{w}_{h_{\text{i}}}\boldsymbol{h}_{t-1}+\boldsymbol{w}_{c_{\text{i}}}\boldsymbol{c}_{t-1}+\boldsymbol{b}_{\text{i}}\right) \tag{5}$$

$$\boldsymbol{f}_t=\sigma\left(\boldsymbol{w}_{x_{\text{f}}}\boldsymbol{x}_t^*+\boldsymbol{w}_{h_{\text{f}}}\boldsymbol{h}_{t-1}+\boldsymbol{w}_{c_{\text{f}}}\boldsymbol{c}_{t-1}+\boldsymbol{b}_{\text{f}}\right) \tag{6}$$

$$\boldsymbol{z}_t=\varphi\left(\boldsymbol{w}_{x_{\text{z}}}\boldsymbol{x}_t^*+\boldsymbol{w}_{h_{\text{z}}}\boldsymbol{h}_{t-1}+\boldsymbol{b}_{\text{c}}\right) \tag{7}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\boldsymbol{z}_t \tag{8}$$

$$\boldsymbol{o}_t=\sigma\left(\boldsymbol{w}_{x_{\text{o}}}\boldsymbol{x}_t^*+\boldsymbol{w}_{h_{\text{o}}}\boldsymbol{h}_{t-1}+\boldsymbol{w}_{c_{\text{o}}}\boldsymbol{c}_t+\boldsymbol{b}_{\text{o}}\right) \tag{9}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\varphi\left(\boldsymbol{c}_t\right) \tag{10}$$

式中：$\boldsymbol{i}_t$、$\boldsymbol{f}_t$ 和 $\boldsymbol{o}_t$ 分别为输入门、遗忘门和输出门；$\boldsymbol{b}_{\text{i}}$、$\boldsymbol{b}_{\text{f}}$、$\boldsymbol{b}_{\text{c}}$、$\boldsymbol{b}_{\text{o}}$ 为偏置项；$\boldsymbol{c}_t$ 为单元记忆；$\boldsymbol{w}$ 为权重矩阵，下标对应输入门、遗忘门和输出门的不同权重矩阵；$\boldsymbol{z}_t$ 为更新状态；σ 为 sigmoid 激活函数；φ 为 tanh 激活函数。对于每个电力词汇向量 $\boldsymbol{x}_t^*$，前向 LSTM 层通过结合电力词汇 $\boldsymbol{x}_1^*$ 到 $\boldsymbol{x}_t^*$ 的上下文信息对 $\boldsymbol{x}_t^*$ 进行编码，标记为 $\overrightarrow{\boldsymbol{h}}_t$。同样地，后向 LSTM 层根据从 $\boldsymbol{x}_\eta^*$ 到 $\boldsymbol{x}_t^*$ 的上下文信息对 $\boldsymbol{x}_t^*$ 进行编码，该信

息被标记为$\overleftarrow{\boldsymbol{h}}_t$。最后，将$\overleftarrow{\boldsymbol{h}}_t$和$\overrightarrow{\boldsymbol{h}}_t$连接起来，代表电力词汇 t 的编码信息，表示为$\boldsymbol{h}_t=\left[\overleftarrow{\boldsymbol{h}}_t,\overrightarrow{\boldsymbol{h}}_t\right]$。

2.3　CRF 层

BiLSTM 层虽然考虑了上下文信息，但没有考虑标签之间的依赖关系，无法保证在每个输出层获得准确的预测结果，因此会出现不符合训练集约束的预测结果，如：一类实体的首字标签后只允许连接同类实体的非首字标签，而不允许直接出现其他类实体非首字标签，具体表现为：电力设备首字标签“B-EQU”后面只能为同类中间或末尾标签“I/E-EQU”或者“O”，而不允许连接其他类实体如生产厂商非首字标签“I-MF”。因此，对于电力领域的实体和关系抽取问题，本文结合配电网运行及巡检数据特征，在 BiLSTM 特征学习的基础上，采用条件随机场(condition random field，CRF)模型来学习约束规则，通过考虑相邻实体标签之间的依赖关系来获取全局的最优标签序列，对 BiLSTM 网络层的输出进行修正，从而减少模型输出非法序列的可能性，弥补 BiLSTM 层缺点，避免出现标签预测不合理的情况，实现电力领域实体和关系的精准抽取。

对于任意输入的电力文本序列 $X=\{x_1,x_2,x_3,\cdots,x_n\}$，假定 $\boldsymbol{P}$ 是 BiLSTM 的输出得分矩阵，$\boldsymbol{P}$ 的大小为 $n\times m$，其中 n 为词的个数，m 为标签个数，矩阵元素 $\boldsymbol{P}_{i',j'}$表示第 i'个词的第 j'个标签的输出分数。引入转移得分方阵 $\boldsymbol{A}$，目的是描述标注序列中相邻 2 个标签之间的转移概率，即对于当前位置的标签和上一个位置的标签，根据它们之间的关系赋予一个得分值，其中矩阵元素 $\boldsymbol{A}_{i',j'}$表示标签 i'转移到标签 j'的转移概率。

对标签序列 $y=\{y_1,y_2,\cdots,y_n\}$，其得分函数为：

$$S(X,y)=\sum_{i=0}^{n}A_{y_{i'}y_{i'+1}}+\sum_{i=1}^{n}P_{i',y_{i'}} \tag{11}$$

得分函数的值越高，给定的输入序列和标签序列同时出现的概率越大。

对所有可能的序列路径进行归一化处理，进一步得到关于预测标签序列 y 的概率计算式，即：

$$P(y|X)=\frac{\exp\left(S(X,y)\right)}{\sum\limits_{\tilde{y}\in Y_X}S(X,\tilde{y})} \tag{12}$$

式中：Y_X 表示所有可能的标记集合，包括不符合标注策略 BIESO 标注规则的序列；$\tilde{y}$ 表示真实正确的标注序列。

在训练过程中，预测序列的似然函数为：

$$\ln\left(P(\tilde{y}|X)\right)=S(X,\tilde{y})-\ln\left(\sum_{\tilde{y}\in Y_X}\mathrm{e}^{S(X,y)}\right) \tag{13}$$

在解码阶段，预测总得分最高的序列作为最优序列输出，即：

$$y^*=\arg\max_{\tilde{y}\in Y_X}S(X,\tilde{y}) \tag{14}$$

式中：y^*为得分最高的电力文本标签序列。

以“变压器”的序列标注过程为例，首先 PowerRoberta 预训练生成的向量在经过 BiLSTM 后得到各个字符的隐藏向量 $\boldsymbol{h}_1$—$\boldsymbol{h}_3$，向量各维度的数值对应各个电力实体标签的得分，在 CRF 层根据每个标签的得分和不同标签之间的转移得分计算综合得分，再将当前综合得分最大的标签路径作为最优标签路径，如图 3 中实线所示。最优标签为“变压器”的最终标注结果，即“B-EQU-H、I-EQU-H、E-EQU-H”，从而实现了对电力文本“变压器”的实体关系识别抽取。

2.4　对抗训练

在配电网运行及巡检数据中，常混杂数字、中英文缩写，甚至可能存在错别字，且非专家标注难免存在错误。这些特殊或不良数据在训练过程中成为噪声干扰，干扰模型的训练学习，导致模型产生置信度较高的错误决策，最终影响传统抽取模型的识别和实际应用。此外，由于电力领域的实体关系抽取数据集规模不够大，模型易识别已出现的词，但较难理解陌生的词，存在过拟合问题，进一步降低了实际模型抽取的准确率。为了解决上述问题，本文在联合抽取过程中增加对抗训练，将扰动添加到模型的嵌入层词向量中作为扰动样本与原始样本一起训练，从而提升模型整体的学习能力和泛化能

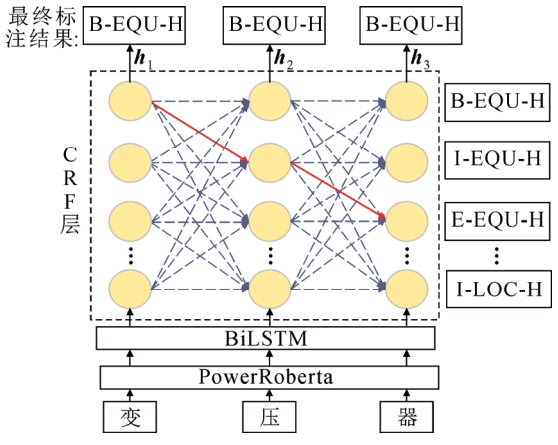


图 3　CRF 标注过程示例

Fig.3　Example of CRF annotation process

力[16-18]，获得更好的训练效果，实现模型在电力领域更高精度的识别和抽取。

然而，对抗训练会引入过大的干扰噪声，可能会破坏数据的有效特征，降低模型性能，因此需要找到合适的噪声水平来平衡模型的性能和鲁棒性。同时，为了避免梯度急速上升出现简单粗暴的“一步到位”现象、导致无法获取约束内最优点，本文采用投影梯度下降(projected gradient descent，PGD)的迭代方法[16]，通过“小步走，多走几步”的迭代策略，在保证一定词向量扰动水平的同时，实现最优点的快速搜寻。把经过 PowerRoberta 处理之后获得的初始向量表示为 $\boldsymbol{E}=[\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3,\cdots, \boldsymbol{E}_n]$，具体计算式为：

$$\boldsymbol{E}_{t+1}=\prod\nolimits_{\boldsymbol{E}+\boldsymbol{U}}\left(\boldsymbol{E}_t+\alpha g\left(\boldsymbol{E}_t\right)/\left\|g\left(\boldsymbol{E}_t\right)\right\|_2\right) \tag{15}$$

$$g\left(\boldsymbol{E}_t\right)=\nabla_{\boldsymbol{E}} L\left(\theta,\boldsymbol{E}_t,\lambda\right) \tag{16}$$

式中：$\boldsymbol{U}=\boldsymbol{r}\in\mathbf{R}^d:\left\|\boldsymbol{r}\right\|_2\le\varepsilon$ 为扰动的约束空间，表示一组选定的扰动，也是无穷范数球的球面限制，限制对抗扰动的幅值；g 为梯度；θ 为模型参数；λ 为样本的真实标签；$L\left(\theta,\boldsymbol{E}_t,\lambda\right)$ 为单个样本的损失值；α 为每一小步的步长。

由式(15)、(16)可以看出，在 $\boldsymbol{U}$ 范围内进行了多步小的对抗训练，对每次训练词嵌入空间的扰动基于上一次扰动不断叠加，然后取最后一次的梯度来更新网络参数。步骤如图 4 所示。

如附录 A 图 A2 所示，本文使用 PowerRoberta 预训练模型获得文本的嵌入向量表示；随后将 PGD 对抗训练策略纳入 PowerRoberta 层，直接对每个嵌入向量组成的参数矩阵进行扰动，计算初始损失值，反向传播得到梯度，进而根据攻击后参数更新的模型权重，计算攻击后的损失值和梯度；然后恢复嵌入层初始权重，累加原始梯度和攻击后的梯度，对模型权重进行更新，以进行对抗训练，并将对抗样本 E_{adv} 输入 BiLSTM 中进行训练，最后经过 CRF 层约束输出最优序列标签。

3 电力知识图谱存储与可视化

本文从改进标注策略、构建针对电力文本的 PowerRoberta 模块、引入对抗训练 3 个方面有效提升了抽取电力实体和关系的准确度，在兼顾抽取实体和关系类型的同时，解决了传统序列标注方法无法抽取重叠实体关系的弊端，以及构建电力知识图谱过程中知识抽取的关键性难题。基于 BIESO+

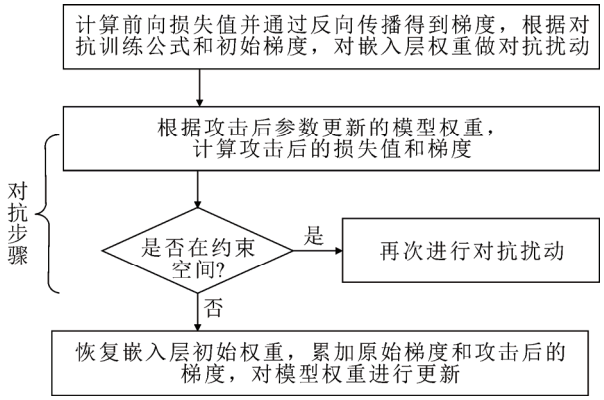


图 4 PGD 对抗流程图

Fig.4 PGD adversarial flowchart

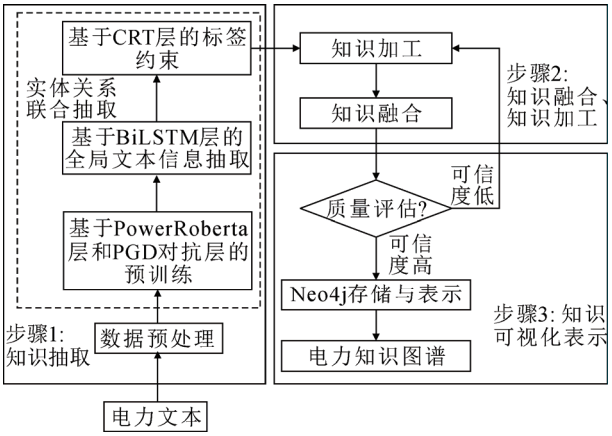


图 5 电力知识图谱构建过程

Fig.5 Construction process of power knowledge graph

Entity+Relation+Role 标注策略和 PowerRoberta-PGD-BiLSTM-CRF 模型抽取，可构建电力知识图谱，其构建流程如图 5 所示。

附录 A 图 A3 基于本文提出的知识图谱构建方法，选取某城市一座典型 220 kV 变电站，检索其全年 7 707 条巡检数据，经过知识抽取后，共形成 574 个节点和 567 个实体间关系，并采用 Neo4j 自动生成可视化电力知识图谱。

4 实验与结果分析

4.1 实验设置

4.1.1 数据集

对于本文提出的电力实体关系联合抽取模型 PowerRoberta-PGD-BiLSTM-CRF，为验证其在电气领域上实体关系联合抽取的有效性，本节面向配电网运行及巡检数据，选取了 2020—2022 年度某城域配电网变电站巡检记录以及设备台账、资产数据库文本信息作为原始语料进行实验测试。

在数据预处理阶段，为了适应电气领域实体和关系的专业性和准确性，将《电力工程基本术语标准》、《电力行业词汇》、《电力分配变压器术语》作为词典对文本语料进行分词处理，将人工标注的7 707 条数据用 sklearn.model_selection.train_test_split 函数以 8:2 的比例划分为训练集和测试集。

4.1.2　评价指标

采用准确率、召回率以及 F_1 值作为实体关系识别结果的评估标准。其中，准确率(precision，P_{Prec})表示识别出的正确的实体关系数量占实际识别出实体关系的总数量的比例，其计算式为：

$$P_{\mathrm{Prec}}=\frac{N_{\mathrm{C}}}{N_{\mathrm{B}}} \tag{17}$$

召回率(recall，R_{Rec})表示识别出的正确的实体关系数量所占理论应该识别出的总数量的比例，其计算式为：

$$R_{\mathrm{Rec}}=\frac{N_{\mathrm{C}}}{N_{\mathrm{A}}} \tag{18}$$

用 F_1 值对准确率和召回率进行综合评价，根据 F_1 值可以对模型的识别效果进行综合评估，其计算式为：

$$F_1=\frac{2P_{\mathrm{Prec}}R_{\mathrm{Rec}}}{P_{\mathrm{Prec}}+R_{\mathrm{Rec}}} \tag{19}$$

式中：N_{A} 表示标注数据集中的实体总数量，即标注数据集给出的实际标准答案；N_{B} 表示模型经过识别得到的实体总数量；N_{C} 表示模型识别出的正确的实体总数量，即模型识别出的实体和实际标注数据集中的实体一致的数量。

4.1.3　实验环境及参数设置

基于 Tensorflow 框架搭建联合抽取模型，具体训练环境配置如下：CPU 为 Inter(R) Core(TM) i7-6700HQ CPU@2.60 GHz，GPU 为 NVIDIA GeForce GTX1060，Python 版本为 3.7，Tensorflow1.14，操作系统为 Windows 10。

具体超参数设定如下：Transformer 层数为 12，隐藏层维数为 768，训练轮次为 10，训练批尺寸为 8，PowerRoberta 读取最大序列长度为 250，优化器为适应性矩估计 Adaptive Moment Estimation(Adam)，初始学习率为 5×10^{-5}。

4.2　实验结果分析

4.2.1　模型抽取性能对比实验

本文首先测试了 PPBC 联合抽取模型对人工电力数据集的抽取性能，并与双向编码器表征法(bidirectional encoder representations from transformers，BERT)流水线模型、Noveltagging 模型[14]、Copyrrl 模型[19]和 Casrel 模型[20]进行对比，结果如图 6 所示。

分析实验数据可知，本文提出的面对电力知识图谱构建的联合抽取方法的抽取准确率、召回率和 F_1 值 2 个评测指标均优于对比算法。比基于 BERT 的流水线方法的 P_{Prec}、R_{Rec}、F_1 值分别提升 8.37%、7.85%、8.10%，比 Noveltagging 模型的 P_{Prec}、R_{Rec}、F_1 值分别提升 19.20%、43.52%、35.38%。与以上 2 个流水线模型的对比结果证明了实体和关系联合抽取方法的优越性：能够有效解决流水线式方法错误传播、信息交互不充分、产生冗余信息的问题，提升了抽取准确率；同时本文改进的标注策略能更有效地应用在实际构建电力知识图谱的知识抽取任务中，其 P_{Prec}、R_{Rec}、F_1 值相较于原始标注策略有明显提升。

进一步对比分析本文模型与基于 Sequence-to-Sequence 方法的 Copyrrl、Casrel 模型性能可知，其 P_{Prec} 分别提升了 13.77%和 3.27%，R_{Rec} 分别提升了 17.97%和 3.74%，F_1 值分别提升了 16.14%和 3.52%。这是因为 Copyrrl 难以准确识别包含大量实体边界信息的长文本，导致抽取准确率不高；而 Casrel 模型虽然能很好地解决重叠关系问题，但回避了对实体关系类型的预测和抽取，在实际构建电力知识图谱的应用中表现不佳。而本文模型具有同步抽取实体和关系类型、抽取准确率高、解决序列标注重叠抽取问题的优势，因此在面向配电网运行及巡检电力文本数据时有更优异的抽取表现，提升了构建知识图谱上游任务知识抽取的准确率。

4.2.2 引入对抗训练对比实验

将本文的 PPBC 联合抽取模型与无对抗训练的 PowerBert-BiLSTM-CRF(PBC)模型和采用快速梯

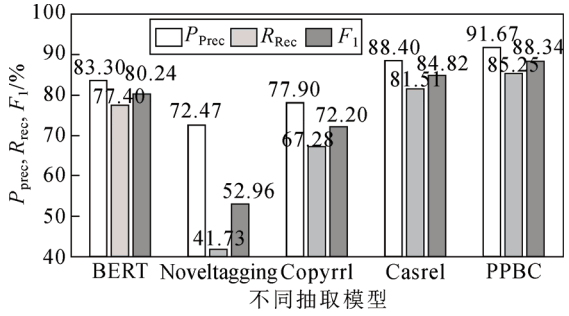


图 6　不同模型抽取性能比较

Fig.6　Comparison of extraction performance of different models

度法(fast gradient method，FGM)的 PowerBert-FGM-BiLSTM-CRF(PFBC)模型进行对比。实验结果表明，加入对抗训练后，模型在抽取在准确率、召回率和 F_1 值上有更好的表现，如图 7 所示。由图 7 可知，加入对抗训练后，无论是 FGM 还是 PGD，其准确率、召回率和 F_1 都显著上升，说明引入对抗训练对实体关系联合抽取准确率的提升效果显著。进一步分析可知，增加 FGM 训练后的 P_{Prec}、R_{Rec}、F_1 值分别提升了 2.29%、5.24%、3.88%，而使用 PGD 训练后的模型 P_{Prec}、R_{Rec}、F_1 值分别提升了 5.23%、6.63%、6%。这是因为对抗训练 FGM 梯度上升的跨步大，存在无法找到约束内的最优点的可能。而本文提出的模型采用 PGD“小步走”策略进行数次迭代，能够快速找到最优解，保证了新模型的识别效果和鲁棒性提升。

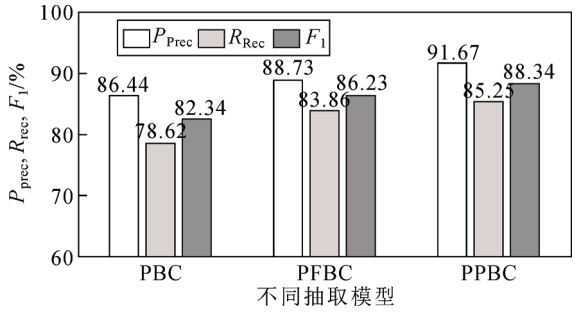


图 7 对抗训练对比图

Fig.7 Comparison of adversarial training

4.2.3 重叠关系实验

为了进一步验证所提出 PPBC 模型在提取重叠关系三元组方面的性能，本文根据关系三元组的不同重叠模式，将标注的电力文本数据集中的句子分成 3 类，即：正常、实体对重叠(entity pair overlap, EPO)和单一实体重叠(single entity overlap, SEO)。对 3 类不同重叠模式的数据分别进行训练。同时，为了更详细地展示模型在处理重叠关系问题中的抽取性能和泛化能力，使用专门评估提取重叠关系三元组模型的国际通用文本数据集纽约时报(New York Times，NYT)和 WebNLG 数据集[21-22]，一起对正常、EPO、SEO 3 类数据进行有效性评测。NYT 数据集包含 56 195 条句子，其中 34.13%的句子存在重叠现象，WebNLG 数据集包含 5 019 条句子，其中 68.2%的句子存在重叠现象。实验结果见图 8。

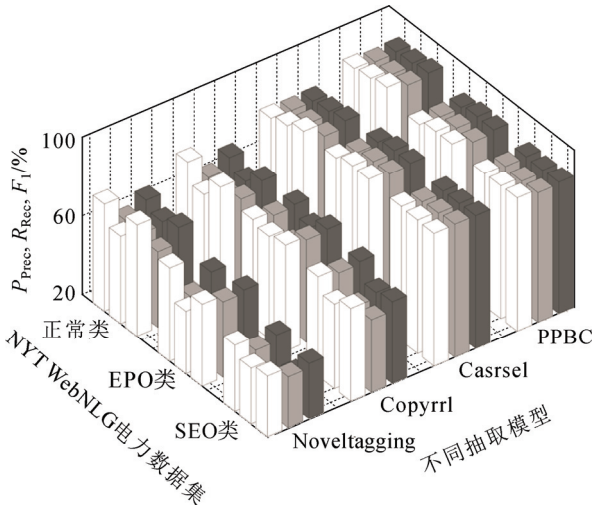


图 8 正常类、EPO 类、SEO 类的 P_{Prec}、R_{Rec}、F_1 值

Fig.8 P_{Prec}, R_{Rec}, F_1 values of normal, EPO class, SEO class

由实验数据可以看出，本文所提出模型在 3 个数据集的正常类表现显著优于其他模型，本文模型在正常类电力数据集上取得的最大 P_{Prec}、R_{Rec}、F_1 值分别为 92.57%、89.2%、90.85%，在正常类 NYT 上取得的最大 P_{Prec}、R_{Rec}、F_1 值分别为 89.31%、88.18%、88.74%，在正常类 WebNLG 上取得的最大 P_{Prec}、R_{Rec}、F_1 值分别为 90.7%、88.1%、89.38%。而对于模型相对难抽取的 EPO 和 SEO 类，本文方法仍然优于其他关系抽取模型，具体分析(以单一实体重叠 SEO 为例说明)为：在 SEO 类电力数据集上，本文模型比序列标注模型 Noveltagging 的 P_{Prec}、R_{Rec}、F_1 值分别提升 37.09%、40.1%、38.71%，比 Copyrrl 的 P_{Prec}、R_{Rec}、F_1 值分别提升 20.4%、28.77%、25%。在 SEO 类 NYT 数据集上，本文模型比序列标注模型 Noveltagging 的 P_{Prec}、R_{Rec}、F_1 值分别提升 33.67%、39.88%、37.02%，比 Copyrrl 的 P_{Prec}、R_{Rec}、F_1 值分别提升 16.6%、18.32%、17.48%。在 SEO 类 WebNLG 数据集上，本文模型比序列标注模型 Noveltagging 的 P_{Prec}、R_{Rec}、F_1 值分别提升 39.16%、56.27%、50.09%，比 Copyrrl 的 P_{Prec}、R_{Rec}、F_1 值分别提升 24.38%、28.03%、26.33%。结果表明本文方法能够解决序列标注中重叠关系的问题。

需要特别注意的是，虽然 Casrel 模型在 EPO 和 SEO 类的数据集上有良好的表现(这是因为该模型倾向于从一个给定的句子中提取尽可能多数量的关系三元组)，但由于回避了对实体和关系类型的预测，Casrel 在第 4.2.1 节实验中完整电力数据集以及第 4.2.3 节实验中 3 个正常类数据集上的表现均劣于本文模型，说明 Casrel 模型无法准确实现面向配电网知识图谱构建的知识抽取。

5　结论

1）针对构建电力知识图谱关键步骤知识抽取中存在的错误传递、信息冗余等问题，以配电网运行及巡检文本为研究对象，提出一种基于改进的序列标注联合抽取方法，搭建了 PowerRoberta-PGD-BiLSTM-CRF 知识抽取模型，具有同步抽取实体和关系类型、抽取准确率高、解决序列标注重叠抽取问题的优势，提升了构建知识图谱上游任务知识抽取的准确率。

2）构建了电力领域专属预训练 PowerRoberta 模型，同时引入 PGD 对抗训练，增加模型的泛化能力，因此本文方法在面向配电网运行及巡检电力文本数据时具有更优异的抽取表现，提升了实体关系联合抽取准确率。

3）基于电力知识数据集和公共文本数据 NYT、WebNLG 进行了实验，结果表明本文提出的方法能够解决序列标注中重叠关系的问题，验证了实体关系联合抽取模型的有效性。通过 Neo4j 图数据库对抽取得到的电力实体关系三元组进行存储，将构建的知识图谱进行可视化表示。研究结果对于提升电力领域知识抽取能力、构建电力知识图谱具有良好的工程应用价值。

附录见本刊网络版(http://hve.epri.sgcc.com.cn)。

参考文献　References

[1] 张　宁，余扬昊，章家维，等. 面向电气工程学科的“大数据技术与应用”课程构建与教学实践[J]. 中国电机工程学报，2022，43(19)：7749-7758.
ZHANG Ning, YU Yanghao, ZHANG Jiawei, et al. Course construction and teaching practice of “big data technology and its applications” curriculum for electrical engineering discipline[J]. Proceedings of the CSEE, 2022, 43(19): 7749-7758.

[2] 高德荃，刘广旭，冯　宝，等. 面向 SDH 光传输系统告警排查的领域知识图谱构建方法[J]. 电力信息与通信技术，2023，21(9)：60-66.
GAO Dequan, LIU Guangxu, FENG Bao, et al. Domain knowledge graph construction method for SDH optical transmission system alarm troubleshooting[J]. Electric Power Information and Communication Technology, 2023, 21(9): 60-66.

[3] 刘　蓓，尚银辉，刘　绚，等. 配电线路跳闸填报文本智能挖掘方法[J]. 高电压技术，2021，47(2)：445-453.
LIU Bei, SHANG Yinhui, LIU Xuan, et al. Inteligent mining method for line trip filling texts in distribution systems[J]. High Voltage Engineering, 2021, 47(2): 445-453.

[4] 刘春梅，郭　岩，俞晓明，等. 针对开源论坛网页的信息抽取研究[J]. 计算机科学与探索，2017，11(1)：114-123.
LIU Chunmei, GUO Yan, YU Xiaoming, et al. Information extraction research aimed at open source web pages[J]. Journal of Frontiers of Computer Science and Technology, 2017, 11(1): 114-123.

[5] 周远翔，葛佳敏，陈健宁，等. 基于巡视文本挖掘的配电设备运行环境风险预测[J]. 高电压技术，2022，48(4)：1213-1225.
ZHOU Yuanxiang, GE Jiamin, CHEN Jianning, et al. Risk prediction for operational environments of power distribution equipment based on inspection text mining[J]. High Voltage Engineering, 2022, 48(4): 1213-1225.

[6] 蒋　晨，王　渊，陈　珉，等. 基于 Kbert 文本聚类模型的变压器故障识别[J]. 高电压技术，2022，48(8)：2991-3000.
JIANG Chen, WANG Yuan, CHEN Min, et al. Transformer fault recognition based on Kbert text clustering model[J]. High Voltage Engineering, 2022, 48(8): 2991-3000.

[7] WANG B L, LU W, WANG W, et al. A neural transition-based model for nested mention recognition[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018: 1011-1017.

[8] 宋　宇，余永忠，梁旭常，等. 基于电能计量管理系统数据分析的配电网线变关系校验方法[J]. 高电压技术，2021，47(12)：4461-4470.
SONG Yu, YU Yongzhong, LIANG Xuchang, et al. Calibration method of line transformer relationship in distribution network based on data analysis of electric energy metering management system[J]. High Voltage Engineering, 2021, 47(12): 4461-4470.

[9] ZHOU G D, SU J, ZHANG J, et al. Exploring various knowledge in relation extraction[C] // Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor, Michigan: ACL, 2005: 427-434.

[10] CHAN Y S, ROTH D. Exploiting syntactico-semantic structures for relation extraction[C] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA: ACL, 2011: 551-560.

[11] 鄂海红，张文静，肖思琪，等. 深度学习实体关系抽取研究综述[J]. 软件学报，2019，30(6)：1793-1818.
E Haihong, ZHANG Wenjing, XIAO Siqi, et al. Survey of entity relationship extraction based on deep learning[J]. Journal of Software, 2019, 30(6): 1793-1818.

[12] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures[C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: ACL, 2016: 1105-1116.

[13] ZHANG M S, ZHANG Y, FU G H. End-to-end neural relation extraction with global optimization[C] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: ACL, 2017: 1730-1740.

[14] ZHENG S C, WANG F, BAO H Y, et al. Joint extraction of entities and relations based on a novel tagging scheme[C] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL, 2017: 1227-1236.

[15] DAI D, XIAO X Y, LYU Y J, et al. Joint extraction of entities and overlapping relations using position-attentive sequence labeling[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI, 2019: 6300-6308.

[16] 任　奎，ZHENG Tianhang，秦　湛，等. 深度学习中的对抗性攻击和防御[J]. 工程(英文)，2020，6(3)：307-339.
REN Kui, ZHENG Tianhang, QIN Zhan, et al. Adversarial attack and defense in deep learning[J]. Engineering, 2020, 6(3): 307-339.

[17] 王道累，孙嘉珺，张天宇，等. 基于改进生成对抗网络的玻璃绝缘子自爆缺陷检测方法[J]. 高电压技术，2022，48(3)：1096-1103.
WANG Daolei, SUN Jiajun, ZHANG Tianyu, et al. Self-explosion defect detection method of glass insulator based on improved generative adversarial network[J]. High Voltage Engineering, 2022, 48(3):

1096-1103.

[18] 李　晖，刘　栋，梁涵卿，等. 基于条件生成对抗网络随机场景的电力系统日前多阶段优化调度[J]. 新型电力系统，2023，1(3)：272-282.
LI Hui, LIU Dong, LIANG Hanqing, et al. Conditional-generative-adversarial-network-based day-ahead multi-stage stochastic scheduling of power systems[J]. New Type Power Systems, 2023, 1(3): 272-282.

[19] ZENG X R, HE S Z, ZENG D J, et al. Learning the extraction order of multiple relational facts in a sentence with reinforcement learning[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: ACL, 2019: 367-377.

[20] WEI Z, SU J, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Washington DC, USA: ACL, 2020: 1476-1488.

[21] ZENG X R, ZENG D J, HE S Z, et al. Extracting relational facts by an end-to-end neural model with copy mechanism[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL, 2018: 506-514.

[22] RIEDEL S, YAO L, MCCALLUM A K. Modeling relations and their mentions without labeled text[C]// Proceedings of ECML PKDD. Barcelona, Spain: Springer, 2010: 148-163.



束嘉伟
2000—，男，硕士生
主要研究方向为电力知识图谱
E-mail: Shujiawei@tju.edu.cn

SHU Jiawei



杨　挺(通信作者)
1979—，男，博士，教授，博导
主要研究方向为电力信息物理融合系统、配用电系统需求侧响应
E-mail: yangting@tju.edu.cn

YANG Ting
Ph.D., Professor
Corresponding author



耿毅男
1993—，男，博士生
主要研究方向为综合能源系统
E-mail: gengyn3@outlook.com

GENG Yinan
Ph.D. candidate



于　洁
1996—，女，博士生
主要研究方向为电力大数据分析，智能电网计量设备误差估计
E-mail: yujie1010@tju.edu.cn

YU Jie
Ph.D. candidate

收稿日期　2023-10-11　修回日期　2024-03-11　　编辑　程子丰

附录 A

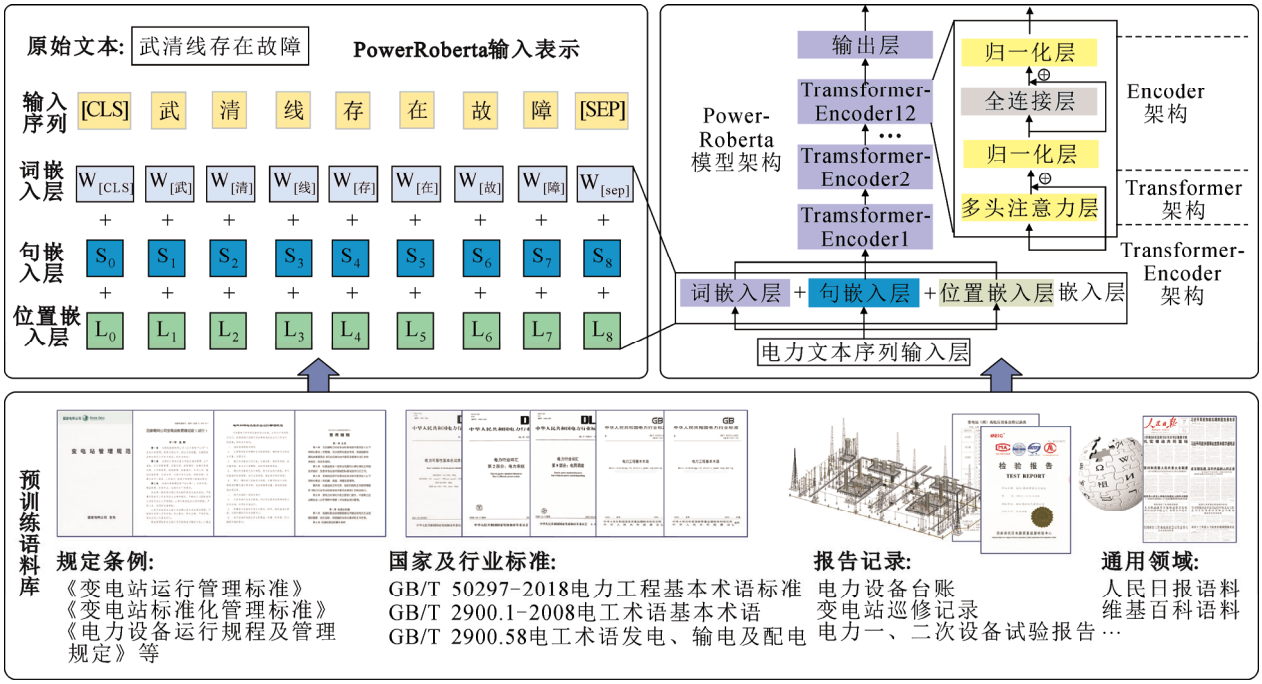


图 A1　PowerRoberta 层架构及输入表示

Fig.A1　PowerRobertsa layer architecture and input representation

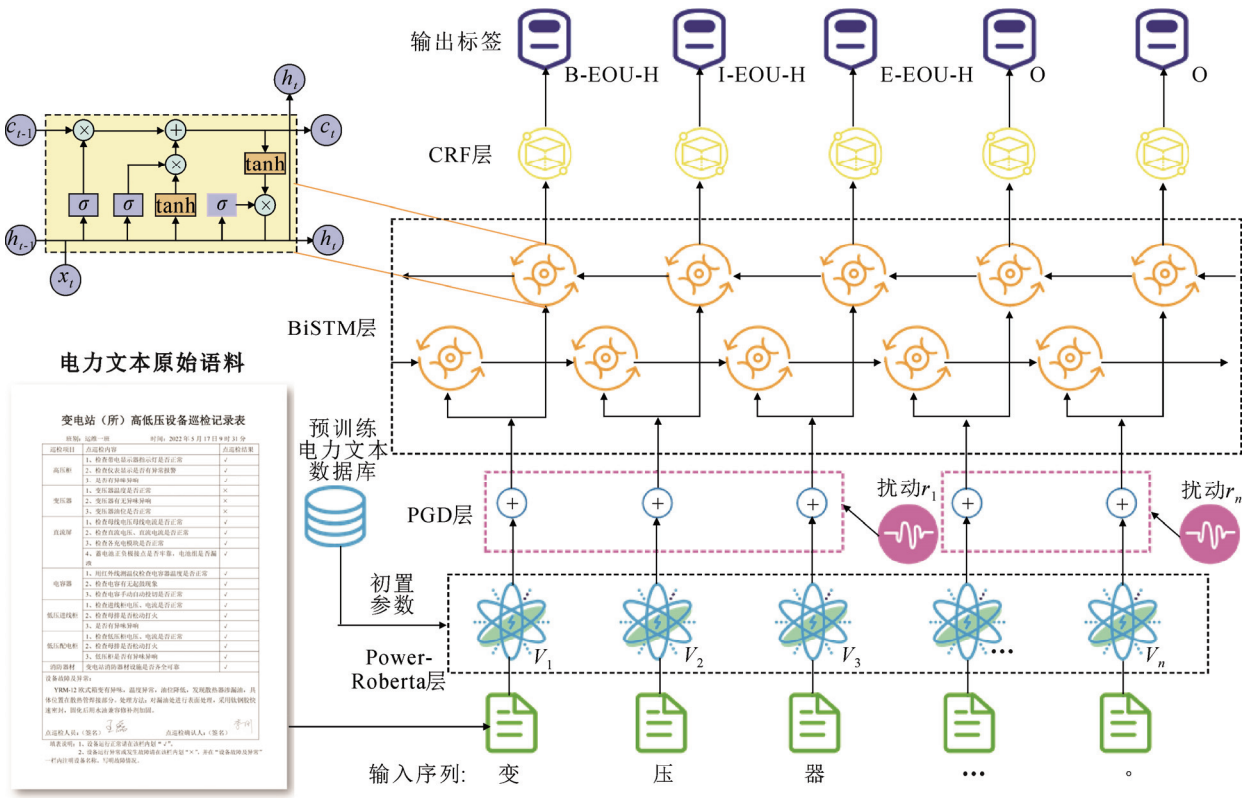


图 A2　PowerRoberta-PGD-BiLSTM-CRF 模型结构图

Fig.A2　PowerRobertsa-PGD-BiLSTM-CRF model structure diagram

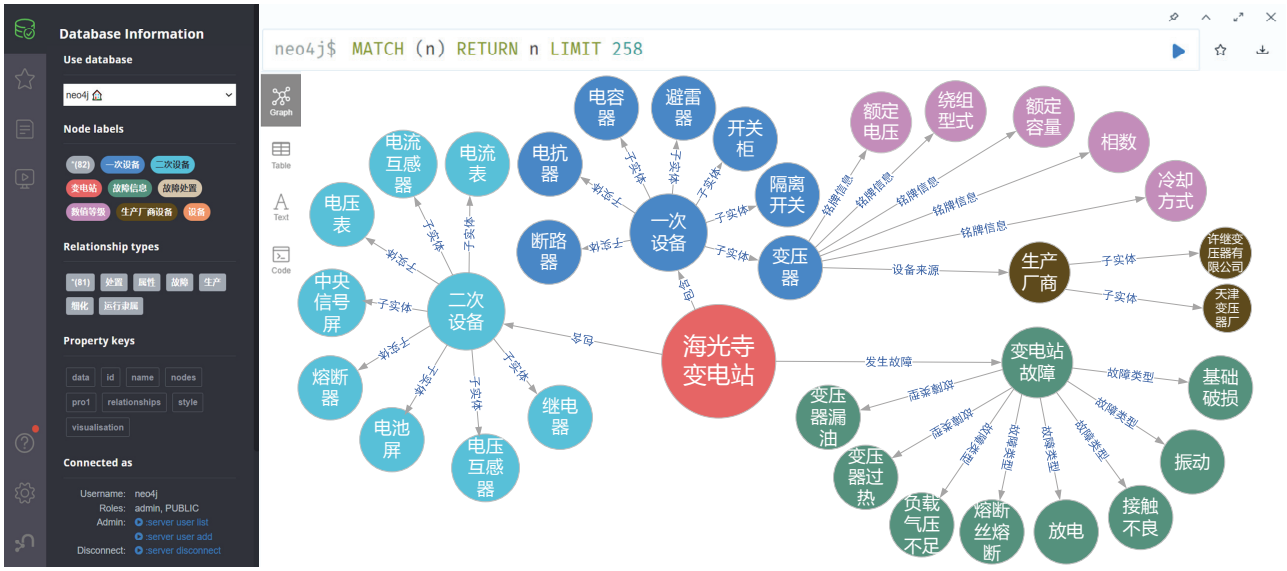


图 A3　电力知识图谱部分示意图

Fig.A3　Schematic diagram of power knowledge graph