DOI:10.12158/j.2096-3203.2021.05.025

电网故障处置预案文本中的命名实体识别研究

江叶峰[1]，孙少华[2,3]，仇晨光[1]，王波[2,3]，戴则梅[2,3,4]，李杰[1]
（1. 国网江苏省电力有限公司，江苏 南京 210024；
2. 南瑞集团（国网电力科学研究院）有限公司，江苏 南京 211106；
3. 国电南瑞科技股份有限公司，江苏 南京 211106；
4. 南瑞集团有限公司智能电网保护和运行控制国家重点实验室，江苏 南京 211106）

摘 要：电网故障处置预案是电网故障处置的重要参考，对电网故障处置预案文本中各类电力设备、名称编号等细粒度的关键实体信息进行抽取，是实现计算机学习理解预案内容并进一步支撑故障处置智能化的重要基础。文中提出一种基于深度学习的电网故障处置预案文本命名实体识别技术，首先采用字向量表征预案文本，然后将注意力机制以及双向长短期记忆网络相结合，有所侧重地提取实体词深层字符特征，最后采用条件随机场求解最优序列化的标注。算例表明：文中所提预案文本命名实体识别模型不依赖人工特征，能够自动高效地提取文本特征，准确识别预案文本中细粒度的实体词，满足预案文本中关键实体信息精确定位和识别的要求。

关键词：电网故障处置预案；命名实体识别；字向量；注意力机制；双向长短期记忆网络；条件随机场
中图分类号：TM732 **文献标志码**：A **文章编号**：2096-3203(2021)05-0177-07

0 引言

电网故障处置预案文本广泛应用于指导设备故障后电网状态监测和故障处理。故障发生后，传统处置方式通过人工查阅预案文本，手动处置故障[1—2]，难以满足故障处置的高效性、及时性。随着电网智能化进程的不断推进，电网故障处置机器人[3—4]应能结合调度系统模型“阅读”文本内容，正确理解故障预案文本，实现相关涉事设备的自我调控。而预案文本由调度员人工编写，每条文本质量参差不齐[5]。因此，对文本进行命名实体识别（named entity recognition，NER），解析预案中关键信息序列，对于提升文本的机器可读性具有重要意义。

近年来，国内外学者针对 NER 任务展开了大量研究，文献[6—7]通过建立领域词典，提升了领域内文本实体词识别能力。文献[8]采用统计分类方法识别实体词。随着词向量技术的发展，专家学者逐渐将神经网络引入 NER 任务中，文献[9]采用拓展卷积神经网络对文本序列建模，关注了文本局部知识与全局信息。文献[10]分别利用循环神经网络（recurrent neural network，RNN）以及长短期记忆网络（long short-term memory，LSTM）标注文本，进一步提升了 NER 效果。在电力领域研究中，针对规范性文本浅层学习，文献[11—12]分别以规范的告警文本、停送电计划为研究对象，参考调度平台数据库匹配关键字符，实现了关键实体词与变量的识别。文献[13]基于专家知识库规则自动生成工作票安全措施。对于非规范性电网缺陷文本的深层挖掘，文献[14—17]基于准确的分词库或者高质量的文本数据，识别近义词或同义词，虽实现了文本分类，但均未详细剖析理解文本信息。预案文本的规范性因人而异，文本匹配显然无法满足实体词识别的要求。因此，预案文本中的关键信息学习识别亟待解决。

文中首先分析了预案文本特征，采用字向量表征文本中汉字，将注意力（attention，ATT）机制引入双向长短期记忆网络（bidirectional long short-term memory，BiLSTM），并结合条件随机场（condition random filed，CRF）提出基于 ATT+BiLSTM+CRF 的电网故障预案文本 NER 方法，实现了文本中涉事电气设备、电气参数词等细粒度的关键实体识别。之后，以 F_1 值为评价指标，对比分析了文中模型与常用 NER 模型的识别效果。实验证明，文中所构建模型对于预案文本具有更强的适用性与鲁棒性。

1 电网故障处置预案文本特点

电网故障处置预案文本是电力调度人员通过离线模拟电网事故，监测故障后薄弱点状态参数信息，并结合电网运行状态人工制定的故障处理方案，既包含电网故障时涉事的电厂、机组等电力设备及其状态参数，也包含设备调控、负荷投切等处

收稿日期：2021-03-05；修回日期：2021-05-13
基金项目：国家重点研发计划资助项目（2017YFB0902600）；国家电网有限公司科技项目（SGJS0000DKJS1700840）

置操作。图 1 为预案文本及 NER 标注示例。

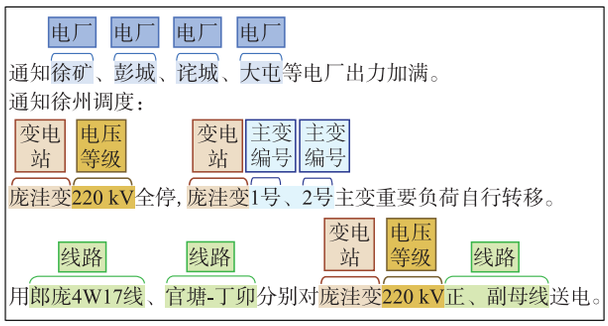

图 1　预案文本及 NER 标注示例

Fig.1　Example of preplan text and NER annotation

根据故障处置的应用需求并结合调度系统中模型划分的实体对象类型，文中将预案文本中电压等级、线路、变电站、开关、母线编号、分区、电厂、机组编号、主变编号作为 NER 的对象，共计 9 类。

不难发现，预案文本中的实体类别是细粒度的。例如，变电站、电厂和分区这 3 类存在很大的相似性，在常规的 NER 任务中通常被粗粒度地划分为“地名”类，然而粗粒度的识别并不适用于电网故障处置的实际需求。

此外，预案文本的表述存在较大的专业性和不规范性，例如电厂名的表述：“华能苏州燃机”“中电滨海风电”等电厂实体词，可细化为所属发电集团、电厂所在地名、电厂类型等复合型电厂实体词；或者在并列表达时仅采用地名代称电厂，如“射阳、彭城等电厂”，其实体词含义及其类别需要结合上下文来分析理解。

同样，线路实体词表述的结构也多种多样，如能量管理系统中标准线路词：“官丁 2569 线”，表述为“官塘-丁卯”“官丁线”。同时线路实体词可能掺杂数字、字母以及用于表达连接的符号“-”、并列的符号“/”等，如“庆安-倪村线”“红柳 4W45/46 双线”等。

综上可以看出，电网故障处置预案文本存在细粒度类别划分、实体词专业性强、语言表述不规范等特点，给 NER 过程带来巨大的困难。

2　基于 ATT+BiLSTM+CRF 的电网故障预案文本 NER

针对电网故障处置预案文本中实体对象的特点，文中采用基于 ATT+BiLSTM+CRF 框架的 NER 方法来进行故障预案文本实体词的辨识与提取。

2.1　数据预处理及文本标注

预案样本数据来自于某电网调度机构 140 个典型故障的预案文本，并根据符号分隔为 4 067 条短句训练样本。通过正则表达式匹配剔除序号、助词等无关词，降低 NER 过程中的噪声。

预案文本中 9 类 NER 对象的标签定义如表 1 所示。对所有预案文本采用“BIO”格式进行标注，标注样例如表 2 所示。

表 1　9 类 NER 对象的标签定义

Table 1　Label definitions of nine NER objects

类别	定义	类别	定义
电压等级	Level	分区	Partition
线路	Line	电厂	Plant
变电站	Substation	机组编号	Set_num
开关	Switch	主变编号	Sub_num
母线编号	Bus_num		

表 2　BIO 标注样例

Table 2　Annotated example of BIO

文本	标注	文本	标注
官	B-Line	9	I-Line
丁	I-Line	单	I-Line
2	I-Line	线	I-Line
5	I-Line	故	O
6	I-Line	障	O

其中，B 为实体词起始词，B-Line 为线路名的起始字；I 为实体词非首字，I-Line 为线路名的非首字；O 为非实体词。

2.2　字向量

文本所用的识别框架需要先将语料中的文字表示成向量形式作为模型的输入，目前学术界主要有 2 种方式：一是词向量形式，将句子切分成多个词，对每个词进行向量化；二是字向量形式，直接将句子中的每个字表示成向量。由于通用领域的分词词典在电力领域适用性较低，会出现明显的分词错误，进而导致模型的性能指标下降，所以文中采用字向量的方式，对语料进行向量化。

字向量化表示的方式有 2 种：One-Hot 方式和分布式方式[18]。但是 One-Hot 方式生成的字向量没有融入任何的语义信息，而且字汇表过大，会造成维度爆炸。分布式方式是将字映射为连续稠密的低维实值向量，较好地解决了 One-Hot 的缺陷问题，所以文中采用分布式方式对字进行向量化。

目前，基于通用语料的预训练模型生成字向量的方式已经在多个通用领域中取得了优异的成绩。但在电力系统领域，由于语料不匹配，效果并不理想。故文中使用目前在 NER 任务中最优的 Bert 预训练模型[19]，在某调度机构的大量相关电力文档上进行训练，得到适用于电力领域的专用预训练模

型,将字映射为 768 维的字向量。

字向量表征的文本可以在模型训练中自动获取文本的字符级特征,从而提升 NER 模型在工程领域文本的适用性和准确率。

2.3 ATT+BiLSTM+CRF 模型

2.3.1 BiLSTM 模块

BiLSTM[20] 是双向结构在 LSTM 上的应用,其每个单元结构与常规 LSTM 的单元相同,只是整体上多了一个按照反方向处理序列的隐层。BiLSTM 模型的结构示意如图 2 所示。

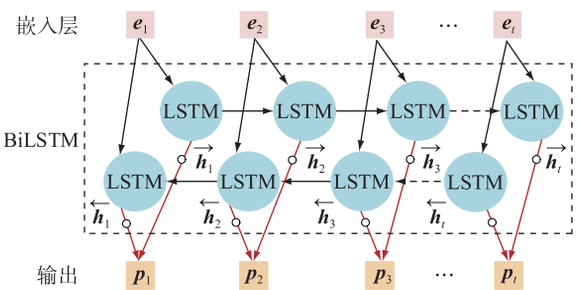

图 2 BiLSTM 结构示意

Fig.2 Schematic diagram of BiLSTM

BiLSTM 的训练方法也采用通过时间的反向传播算法,其前向与反向传播的过程与常规的 RNN 类似。在 t 时刻,BiLSTM 的正向计算过程一般遵循下式:

$$\sigma(x) = 1/(1 + e^{-x}) \tag{1}$$

$$\tanh x = (e^{x} - e^{-x})/(e^{x} + e^{-x}) \tag{2}$$

$$\overrightarrow{\boldsymbol{i}}(t) = \sigma(\boldsymbol{W}_{ei}\boldsymbol{e}_t + \boldsymbol{W}_{hi}\overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{W}_{ci}\overrightarrow{\boldsymbol{c}}_{t-1} + \overrightarrow{\boldsymbol{b}}_i) \tag{3}$$

$$\overrightarrow{\boldsymbol{f}}(t) = \sigma(\boldsymbol{W}_{ef}\boldsymbol{e}_t + \boldsymbol{W}_{hf}\overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{W}_{cf}\overrightarrow{\boldsymbol{c}}_{t-1} + \overrightarrow{\boldsymbol{b}}_f) \tag{4}$$

$$\overrightarrow{\boldsymbol{c}}_t = \overrightarrow{\boldsymbol{f}}_t\overrightarrow{\boldsymbol{c}}_{t-1} + \overrightarrow{\boldsymbol{i}}_t\tanh(\boldsymbol{W}_{xc}\boldsymbol{e}_t + \boldsymbol{W}_{hc}\overrightarrow{\boldsymbol{h}}_{t-1} + \overrightarrow{\boldsymbol{b}}_c) \tag{5}$$

$$\overrightarrow{\boldsymbol{o}}(t) = \sigma(\boldsymbol{W}_{eo}\boldsymbol{e}_t + \boldsymbol{W}_{ho}\overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{W}_{co}\overrightarrow{\boldsymbol{c}}_{t-1} + \overrightarrow{\boldsymbol{b}}_o) \tag{6}$$

$$\overrightarrow{\boldsymbol{h}}_t = \overrightarrow{\boldsymbol{o}}_t\tanh\overrightarrow{\boldsymbol{c}}_t \tag{7}$$

式中:$\boldsymbol{e}_t$ 为 t 时刻的模型输入;$\overrightarrow{\boldsymbol{i}}_t$,$\overrightarrow{\boldsymbol{f}}_t$,$\overrightarrow{\boldsymbol{o}}_t$ 分别为 t 时刻正向网络输入门、遗忘门和输出门的输出;$\overrightarrow{\boldsymbol{c}}_t$,$\overrightarrow{\boldsymbol{h}}_t$ 分别为 t 时刻正向网络的细胞状态、隐状态;$\overrightarrow{\boldsymbol{c}}_{t-1}$,$\overrightarrow{\boldsymbol{h}}_{t-1}$ 分别为 $t-1$ 时刻正向网络的细胞状态、隐状态;$\boldsymbol{W}$ 为各自对应的权重矩阵;$\boldsymbol{b}$ 为各自对应的偏置量。

反向网络的计算与正向网络相似,只是 t 时刻反向网络的隐状态 $\overleftarrow{\boldsymbol{h}}_t$ 的计算依赖于 $t+1$ 时刻的隐状态 $\overleftarrow{\boldsymbol{h}}_{t+1}$。

将正反向网络输出的隐状态按位置拼接,得到 $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}}_t \oplus \overleftarrow{\boldsymbol{h}}_t]$,各时刻输出构成完整的隐状态序列 $\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$,$n$ 为输入字数,再通过线性计算,将隐状态向量映射到 k 维,k 为标注集的标签数,从而得到 BiLSTM 网络自动提取的句子特征,记作 $\boldsymbol{P} = \{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_n\} \in \boldsymbol{R}_{n\times k}$($\boldsymbol{R}$ 为大小为 $n \times k$ 的向量空间),为 BiLSTM 最终输出。

对于序列化文本数据,BiLSTM 通过引入门级控制调控序列数据传输,选择性丢弃和保存前序与后序数据,用以更新神经元,有效地解决了 RNN 中长文本梯度弥散的问题。

2.3.2 CRF 模块

CRF 能够通过考虑相邻标签的关系获得一个全局最优的标签序列[21]。对于文本 $\boldsymbol{S}$ 的字向量序列 $\boldsymbol{S} = \{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n\}$,设 $\tilde{\boldsymbol{y}} = \{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$ 为与 $\boldsymbol{S}$ 对应的预测序列,其中每个标签 $\boldsymbol{y}_i \in [1,k]$ 为标注集中类别的索引编号,共 k^n 种可能的序列组合。$\tilde{\boldsymbol{y}}^* = \{\boldsymbol{y}_1^*,\boldsymbol{y}_2^*,\cdots,\boldsymbol{y}_n^*\}$ 表示与 $\boldsymbol{S}$ 对应的标注序列。

定义每一种预测序列的得分如式(8)所示。

$$s(\boldsymbol{S},\tilde{\boldsymbol{y}}) = \sum_{t=0}^{n} A_{\boldsymbol{y}_t,\boldsymbol{y}_{t+1}} + \sum_{t=1}^{n} P_{\boldsymbol{p}_t,\boldsymbol{y}_t} \tag{8}$$

式中:$\boldsymbol{A}$ 为 $(k+2)\times(k+2)$ 的矩阵,加 2 是为了提升鲁棒性,在句子首尾添加了起始状态和终止状态;$A_{\boldsymbol{y}_t,\boldsymbol{y}_{t+1}}$ 为类别 $\boldsymbol{y}_t$ 到 $\boldsymbol{y}_{t+1}$ 的转移得分,代表了实体各标签类别之间的依赖关系;$\boldsymbol{y}_0$,$\boldsymbol{y}_{n+1}$ 分别为起始状态和终止状态;$\boldsymbol{P}$ 为 BiLSTM 的输出矩阵;$P_{\boldsymbol{p}_t,\boldsymbol{y}_t}$ 为第 t 时刻的输出向量 $\boldsymbol{p}_t$ 与 $\boldsymbol{y}_t$ 类对应的得分。

定义标注序列 $\tilde{\boldsymbol{y}}^*$ 在所有可能的序列集合 $\tilde{Y}$ 中的 softmax 概率得分如式(9)所示。

$$p(\tilde{\boldsymbol{y}}^* \mid \boldsymbol{S}) = e^{s(\boldsymbol{S},\tilde{\boldsymbol{y}}^*)} \Big/ \sum_{\tilde{\boldsymbol{y}}\in\tilde{Y}} e^{s(\boldsymbol{S},\tilde{\boldsymbol{y}}^*)} \tag{9}$$

模型训练的目标是使 $p(\tilde{\boldsymbol{y}}^* \mid \boldsymbol{S})$ 趋近于 1。与 BiLSTM 一同训练,不断优化 $\boldsymbol{A}$ 和 BiLSTM 的各参数矩阵。

训练好的模型对测试数据进行预测即可得到最佳标签序列,其计算公式如下:

$$\tilde{\boldsymbol{y}}_m = \operatorname{argmax} s(\boldsymbol{S},\tilde{\boldsymbol{y}}) \tag{10}$$

目前,BiLSTM+CRF 模型在 NER 领域已经取得了广泛的应用,在网络开源语料数据集上也取得了领先的识别效果。然而电网故障处置预案文本与一般性文本存在巨大差异,具有很强的专业性,BiLSTM+CRF 模型难以取得理想的识别效果。文中针对目前 BiLSTM+CRF 模型在电网故障处置预案文本上识别效果的不足,提出一种引入 ATT 机制的 ATT+BiLSTM+CRF 模型。通过在电网故障处置预案文本 NER 中对实体词关键部分分配较多的注意

力，从而提升电网故障处置预案的 NER 效果。

2.3.3 ATT 机制

预案文本的部分内容具有关联性的特征，例如："在徐州西分区进行事故拉限电"，其中"徐州西分区"的字符间关联性更高，"徐"和"在"字的关联性很弱，这说明对于识别文本中的命名实体，每个字符的影响程度不同，在数学中表示为分配的权重不同。因此，文中在 BiLSTM 计算过程中引入 ATT 机制[22]。

注意力模型对 BiLSTM 的输出特征向量序列 $\boldsymbol{P}$ 进行处理，对每个特征向量赋予不同大小的权重，相加后产生新的特征向量，包含文本全局和局部特征。

注意力模型的当前状态 $\boldsymbol{c}_t$ 由 $\boldsymbol{P}$ 中的所有特征向量加权后得到，计算如下：

$$\boldsymbol{c}_t = \sum_{j=1}^{n} (\alpha_{tj} \boldsymbol{p}_j) \tag{11}$$

式(11)中特征向量分配权重 α_{tj} 通过式(12)和式(13)计算得到。

$$\alpha_{tj} = \mathrm{e}^{e_{tj}} \Big/ \sum_{k=1}^{n} \mathrm{e}^{e_{tk}} \tag{12}$$

$$e_{tj} = a(\boldsymbol{c}_{t-1}, \boldsymbol{p}_j) = \boldsymbol{v}_a^n \tanh(\boldsymbol{w}_a \boldsymbol{c}_{t-1} + \boldsymbol{u}_a \boldsymbol{p}_j) \tag{13}$$

式中：e_{tj} 为关联能量，用于量化时刻 j 的输入和时刻 t 的输出之间的关系，由前一时刻的注意力模型输出状态 $\boldsymbol{c}_{t-1}$ 和时刻 j 上的特征向量 $\boldsymbol{p}_j$ 决定，通过函数 a 计算；$\boldsymbol{w}_a$ 为 $\boldsymbol{c}_{t-1}$ 的权值矩阵；$\boldsymbol{u}_a$ 为 $\boldsymbol{p}_j$ 的权值矩阵；$\boldsymbol{v}_a^n$ 为整体的权值矩阵。模型通过训练迭代更新这 3 个矩阵的参数，实现注意力权重的自动分配。

序列 $\tilde{\boldsymbol{c}} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_n\}$ 即为 ATT 机制输出，以此代替 BiLSTM 的输出序列作为 CRF 模块的输入。

2.3.4 ATT+BiLSTM+CRF 模型框架

引入 ATT 机制后的模型整体框架如图 3 所示。

利用 BiLSTM 对文本向量 $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \cdots, \boldsymbol{e}_n\}$ 进行特征提取。对于 t 时刻输入字向量 $\boldsymbol{e}_t$，通过顺序(由左至右序列)输入，计及前文信息与当前时刻信息得到隐状态输出 $\overrightarrow{\boldsymbol{h}}_t$，通过倒序(由右至左序列)输入，计及后文信息与当前时刻信息得到隐状态输出 $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}}_t \oplus \overleftarrow{\boldsymbol{h}}_t]$，隐状态序列经线性映射后得到 $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n\}$ 为 BiLSTM 的输出。

通过 ATT 机制对 $\boldsymbol{P}$ 中特征向量分别进行加权计算得到 $\tilde{\boldsymbol{c}} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_n\}$，有侧重地关注不同的字向量，之后将 $\tilde{\boldsymbol{c}}$ 传递给 CRF 层。

CRF 模型计及标签之间的约束以及相关性，在所有备选标签序列中求得标签序列的最优解。最终识别"官丁 2569 单线"属于线路"Line"。

文中模型训练相关参数设置为：优化器为 Adam；学习率取 10^{-4}；在 BiLSTM 两端增加比例为 0.2 的 Dropout；最大迭代次数限制在 100 000 次；最大容忍次数 earlystop 设为 5；批处理大小为 50。

3 实验分析

3.1 实验数据和评价指标

文中研究的实验环境为 Intel(R) Core i7-8700 CPU 3.2 GHz 处理器，16 GB 内存，GPU NVIDIA 1080Ti，Windows10 操作系统。字向量与训练语料为某电力调度机构的各类电力工作文档，选择开源 Bert 模型作为输入的分布式表示模型；BiLSTM 网络由 Tensorflow 实现。NER 语料为某电网故障处置预案 2015 年的历史版本，文本已分句并经人工标注，共计 5 230 条故障预案例句，按 8∶1∶1的比例分为训练集、验证集、测试集。

电网故障处置预案中文 NER 的评价指标采用综合考虑查准率、查全率的 F_1 测量值。

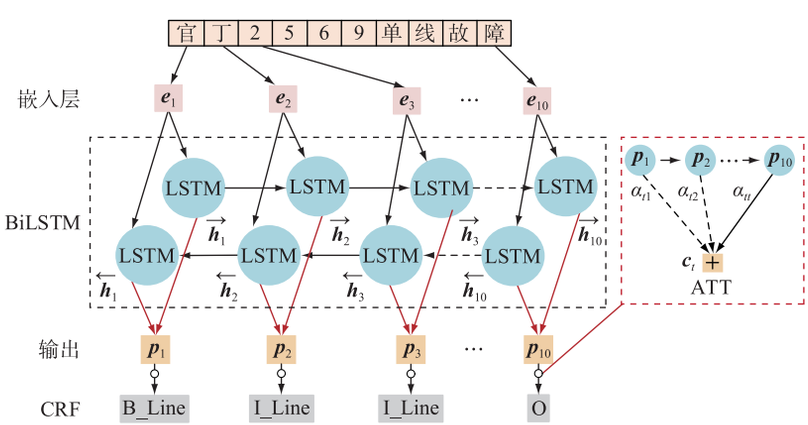

图 3 基于 ATT+BiLSTM+CRF 的 NER 流程

Fig.3 The process of NER based on ATT+BiLSTM+CRF

(1) 查准率。查准率 P 计算公式为:

$$P = T_P/(T_P + F_P) \tag{14}$$

式中:T_P 为正确识别为正样本的实体词数量;F_P 为错误识别为正样本的实体词数量。

(2) 查全率。查全率 R 计算公式为:

$$R = T_P/(T_P + F_N) \tag{15}$$

式中:F_N 为正样本中识别错误的实体词数量。

(3) F_1 测量值。F_1 值计算公式为:

$$F_1 = 2T_P/(2T_P + F_P + F_N) \tag{16}$$

F_1 值综合考虑了查全率与查准率,能够更加全面地分析分类效果。

3.2 不同模型实验设计及性能对比

为了验证文中提出的故障预案文本 NER 框架的优越性,分别设计了 6 组实验。6 组实验使用了相同的电网故障处置预案命名实体语料、字向量输入。实验 1 为基于 BiLSTM 的模型;实验 2 为基于人工特征提取的正则表达式添加 CRF 作用的模型;实验 3 为将实验 2 中的人工特征替换为 RNN 的 RNN+CRF 模型;实验 4 为将 RNN 替换为 LSTM 的 LSTM+CRF 模型;实验 5 为 BiLSTM+CRF 模型,实验 6 为文中模型即基于 ATT+LSTM+CRF 模型。

6 组实验中的不同模型分别对电网故障处置预案中的 9 类实体词进行识别,获得的综合评价指标 F_1 记录值见表 3。

表 3 各模型 F_1 记录表

Table 3 Record chart of F_1 of each model %

模型名称	类别								
	线路	电压等级	母线编号	主变编号	机组编号	开关	分区	电厂	变电站
BiLSTM	76.9	89.5	89.2	92.5	80.2	83.7	84.2	83.8	80.0
人工特征+CRF	86.0	96.7	100	100	90.9	90.9	94.3	88.2	87.9
RNN+CRF	80.1	92.4	94.4	96.0	85.6	86.4	90.5	90.1	85.5
LSTM+CRF	83.8	93.7	96.9	97.2	86.3	88.1	91.6	91.0	87.5
BiLSTM+CRF	90.5	97.6	100	100	91.4	92.8	95.2	92.9	92.4
文中模型	92.7	98.8	100	100	92.8	92.7	96.8	95.8	97.5

由表 3 分析可知:

(1) LSTM+CRF 模型的 NER 实验效果优于 RNN+CRF 模型,其中线路、变电站、电厂实体词识别效果值提高了 3.7%,2.0%,0.9%,但是两者对线路、电压等级、主变编号等类别的识别效果未能超过基于人工特征提取的 CRF 模型,没有很好地体现自动特征提取的优势。而 BiLSTM+CRF 模型在识别线路、变电站、电厂等类别的实体词过程中较 LSTM+CRF 网络模型表现更加优异,F_1 值分别提升了 6.7%,4.9%,1.9%,并且实现了对人工特征+CRF 模型的超越。这是因为 BiLSTM 同时考虑了前序和后序内容,结合上下文语义信息更加充分地提取了文本字符特征。

(2) 单独的 BiLSTM 模型取得的识别效果最差,而结合 CRF 模型后,识别效果显著提升,其中线路实体词的识别效果提升最高,F_1 值提高了13.6%,变电站、机组编号分别提升了 12.4%,11.2%。结合具体文本内容分析,这是由于 CRF 模型计及了相邻标签关联性约束,从而能够更好地识别线路词中的长距离实体词。

(3) BiLSTM+CRF 模型在电网故障处置预案文本中确实可以取得较好的识别效果,F_1 值可以达到 90%以上,而引入了 ATT 机制后,实体词识别整体效果进一步提升,模型识别线路、电厂、变电站的 F_1 值分别提升了 2.2%,2.9%,5.1%,更加符合电网处置预案的识别要求。

4 结语

文中针对电网故障处置预案文本中关键信息辨识的任务,搭建了基于 ATT+BiLSTM+CRF 的电网故障处置预案文本 NER 模型,实现了故障处置预案文本关键信息的 NER。

通过采用字向量特征表征文本,规避了专业领域词向量训练对于人工的依赖以及专业领域词向量表达能力差的缺陷。同时采用字向量可以更好地识别“官塘-丁卯 2569 单线”“官丁线”“官塘-丁卯”“官丁 2569”等不同表述形式的线路实体词,提升了模型对于含复杂实体词电力文本的适用性。

基于 ATT+BiLSTM+CRF 模型可以综合考虑电网故障处置预案文本中的实体词长度较长,并列实体词的简写表达随意性大以及文本长距离造成信息丢失的问题,通过引入 ATT 机制以及 BiLSTM,有所侧重地、自动地学习获取文本特征信息,降低了人工成本,提升了模型的泛化能力。算例表明文中所提模型可以满足电网故障处置预案文本的 NER 任务要求,为电力文本的 NER 提供有效路径。

通过故障预案文本中实体词序列准确识别,文本内容即可实现准确切分和词义理解,进而简化了文本句法结构和语义分析,为机器学习非结构化故障预案文本,搭建电力故障处置预案垂直知识图谱打下重要基础。

参考文献:

[1] 范士雄,李立新,王松岩,等. 人工智能技术在电网调控中的应用研究[J]. 电网技术,2020,44(2):401-411.
FAN Shixiong, LI Lixin, WANG Songyan, et al. Application a-

nalysis and exploration of artificial intelligence technology in power grid dispatch and control[J]. Power System Technology, 2020,44(2):401-411.

[2] 闪鑫,陆晓,翟明玉,等. 人工智能应用于电网调控的关键技术分析[J]. 电力系统自动化,2019,43(1):49-57.
SHAN Xin,LU Xiao,ZHAI Mingyu,et al. Analysis of key technologies for artificial intelligence applied to power grid dispatch and control[J]. Automation of Electric Power Systems,2019,43(1):49-57.

[3] SHAN X,ZHU B Q,WANG B,et al. Research on deep learning based dispatching fault disposal robot technology[C]//2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2). Beijing,China. IEEE,2018:1-6.

[4] 李明节,陶洪铸,许洪强,等. 电网调控领域人工智能技术框架与应用展望[J]. 电网技术,2020,44(2):393-400.
LI Mingjie,TAO Hongzhu,XU Hongqiang,et al. The technical framework and application prospect of artificial intelligence application in the field of power grid dispatching and control[J]. Power System Technology,2020,44(2):393-400.

[5] 邵冠宇,王慧芳,何奔腾. 电网设备缺陷文本的质量评价与提升方法[J]. 电网技术,2019,43(4):1472-1479.
SHAO Guanyu,WANG Huifang,HE Benteng. Quality assessment and improvement method for power grid equipment defect text[J]. Power System Technology,2019,43(4):1472-1479.

[6] 冯艳红,于红,孙庚,等. 基于词向量和条件随机场的领域术语识别方法[J]. 计算机应用,2016,36(11):3146-3151.
FENG Yanhong,YU Hong,SUN Geng,et al. Domain-specific term recognition method based on word embedding and conditional random field[J]. Journal of Computer Applications,2016,36(11):3146-3151.

[7] 郁圣卫,卢奇,陈文亮. 基于领域情感词典特征表示的细粒度意见挖掘[J]. 中文信息学报,2019,33(2):112-121.
YU Shengwei,LU Qi,CHEN Wenliang. Fine-grained opinion mining based on feature representation of domain sentiment lexicon[J]. Journal of Chinese Information Processing,2019,33(2):112-121.

[8] 樊梦佳,段东圣,杜翠兰,等. 统计与规则相融合的领域术语抽取算法[J]. 计算机应用研究,2016,33(8):2282-2285,2306.
FAN Mengjia,DUAN Dongsheng,DU Cuilan,et al. Domain-specific terms extraction algorithm based on combination of statistics and rules[J]. Application Research of Computers,2016,33(8):2282-2285,2306.

[9] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[C]//ICLR. 2016.

[10] COLLOBERT R,WESTON J,BOTTOU L,et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research,2011.

[11] 张旭,魏娟,赵冬梅,等. 一种用于电网故障诊断的遥信信息解析方法[J]. 中国电机工程学报,2014,34(22):3824-3833.
ZHANG Xu, WEI Juan, ZHAO Dongmei, et al. Ananalytic alarm information method for power grid fault diagnosis[J]. Proceedings of the CSEE,2014,34(22):3824-3833.

[12] 佟佳弘,武志刚,管霖,等. 电力调度文本的自然语言理解与解析技术及应用[J]. 电网技术,2020,44(11):4148-4156.
TONG Jiahong, WU Zhigang, GUAN Lin, et al. Power dispatching text analysis and application based on natural language understanding[J]. Power System Technology,2020,44(11):4148-4156.

[13] 曾厉,虞晨曦,刘海艳. 基于智能成票规则及算法的电气工作票系统[J]. 电工技术,2019(24):105-110.
ZENG Li,YU Chenxi,LIU Haiyan. Electrical work order system based on intelligent ticketing rules and algorithms[J]. Electric Engineering,2019(24):105-110.

[14] 曹靖,陈陆燊,邱剑,等. 基于语义框架的电网缺陷文本挖掘技术及其应用[J]. 电网技术,2017,41(2):637-643.
CAO Jing,CHEN Lushen,QIU Jian,et al. Semantic framework-based defect text mining technique and application in power grid[J]. Power System Technology,2017,41(2):637-643.

[15] 邱剑,王慧芳,应高亮,等. 文本信息挖掘技术及其在断路器全寿命状态评价中的应用[J]. 电力系统自动化,2016,40(6):107-112,118.
QIU Jian,WANG Huifang,YING Gaoliang,et al. Text mining technique and application of lifecycle condition assessment for circuit breaker[J]. Automation of Electric Power Systems,2016,40(6):107-112,118.

[16] 刘梓权,王慧芳,曹靖,等. 基于卷积神经网络的电力设备缺陷文本分类模型研究[J]. 电网技术,2018,42(2):644-651.
LIU Ziquan,WANG Huifang,CAO Jing,et al. A classification model of power equipment defect texts based on convolutional neural network[J]. Power System Technology,2018,42(2):644-651.

[17] 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):4162-4172.
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):4162-4172.

[18] 温潇. 分布式表示与组合模型在中文自然语言处理中的应用[D]. 南京:东南大学,2016.
WEN Xiao. Application of distributed representation and composition model in Chinese natural language processing[D]. Nanjing:Southeast University,2016.

[19] DEVLIN J,CHANG M W,LEE K,et al. Bert:pre-training of deep bidirectional transformers for language understanding[J]. Computer Science,2018.

[20] MA J,GANCHEV K,WEISS D. State-of-the-art Chinese word segmentation with Bi-LSTMs[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels,Belgium. Stroudsburg,PA,USA:Association

for Computational Linguistics,2018: 4902-4908.

[21] 陈伟,吴友政,陈文亮,等. 基于 BiLSTM-CRF 的关键词自动抽取[J]. 计算机科学,2018,45(S1):91-96,113.
CHEN Wei,WU Youzheng,CHEN Wenliang,et al. Automatic keyword extraction based on BiLSTM-CRF[J]. Computer Science,2018,45(S1):91-96,113.

[22] BAHDANAU D,CHO K,BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. Computer Science,2014.

作者简介:

江叶峰

江叶峰(1976),男,硕士,高级工程师,从事电网调度运行管理及在线安全稳定分析应用相关工作(E-mail:jyf1976@sina.cn);

孙少华(1992),男,硕士,工程师,从事电力文本挖掘工作;

仇晨光(1977),男,硕士,高级工程师,从事电网调度运行管理及智能调度工作。

Named entity recognition in power fault disposal preplan text

JIANG Yefeng[1], SUN Shaohua[2,3], QIU Chenguang[1], WANG Bo[2,3], DAI Zemei[2,3,4], LI Jie[1]

(1. State Grid Jiangsu Electric Power Co.,Ltd.,Nanjing 210024,China;
2. NARI Group (State Grid Electric Power Research Institute) Co.,Ltd.,Nanjing 211106,China;
3. NARI Technology Co.,Ltd.,Nanjing 211106,China;
4. State Key Laboratory of Smart Grid Protection and Operation Control,NARI Group Co.,Ltd.,Nanjing 211106,China)

Abstract: Power grid fault disposal preplan is an important reference for power grid fault disposal. Hence, extracting fine-grained key entity information such as power equipments, name and number from the preplan is an important basis for the computer to understand the content and further support the intelligent disposal. A named entity recognition technology for power grid fault disposal preplan is proposed based on deep learning. Firstly, the word vector is used to represent the preplan text. Then the word vector features are extracted by combining the attention mechanism and the bidirectional long short-term memory network. Finally, the optimal serialization annotation is solved by the conditional random field. The example shows that the proposed entity recognition model can automatically and efficiently extract text features, thus accurately identifying entity words in the preplan. It proves that the model meets the requirement of extracting key entity information in the preplan better than another commonly used model dose.

Keywords: power grid fault disposal preplan text; named entity recognition; word vector; attention mechanism; bidirectional long short-term memory network; conditional random field

(编辑 陆海霞)