计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(6):210−216 [doi: 10.15888/j.cnki.csa.008525]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向电力规章制度的命名实体识别①

陈　鹏[1], 蔡　冰[1], 何晓勇[2], 金兆轩[2], 金志刚[2], 侯　瑞[3,4]

[1](国网宁夏电力有限公司, 银川 750001)
[2](天津大学 电气自动化与信息工程学院, 天津 300072)
[3](华北电力大学 苏州研究院, 苏州 215123)
[4](华北电力大学 经济与管理学院, 北京 102206)
通信作者: 陈　鹏, E-mail: chenpengtougao2021@163.com

摘　要: 在电力生产的过程中, 往往会产生大量电力相关的文本数据, 但这些数据大多是非结构化数据且体量庞大繁杂, 实现对电力相关数据有效的组织管理可以促进电力企业实现数字资产商品化, 以此为电力企业发掘新的利润增长点. 本文针对将电力行业中的相关规章制度文本进行结构化处理这一问题, 提出了基于字符和二元词组特征的命名实体识别的模型. 在该模型中, 通过使用融合多特征的 BERT 预训练语言模型得到词嵌入表示, 并使用引入相对位置编码的 Transformer 模型和条件随机场作为编码层和解码层, 本文提出的模型在实体类型识别的准确率为 92.64%, 取得了有效的识别效果.

关键词: 命名实体识别; BERT 模型; Transformer 模型; 条件随机场

引用格式: 陈鹏,蔡冰,何晓勇,金兆轩,金志刚,侯瑞.面向电力规章制度的命名实体识别.计算机系统应用,2022,31(6):210–216. http://www.c-s-a.org.cn/1003-3254/8525.html

Named Entity Recognition for Electric Power Regulations

CHEN Peng[1], CAI Bing[1], HE Xiao-Yong[2], JIN Zhao-Xuan[2], JIN Zhi-Gang[2], HOU Rui[3,4]

[1](State Grid Ningxia Electric Power Co. Ltd., Yinchuan 750001, China)
[2](School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)
[3](Suzhou Institute, North China Electric Power University, Suzhou 215123, China)
[4](School of Economics and Management, North China Electric Power University, Beijing 102206, China)

Abstract: In the process of power production, a large amount of power-related text data is often generated, and most of these data are unstructured and large in size. Thus, achieving effective organization and management of these data can promote power companies to produce digital asset products, which can help discover new profit growth points for power companies. Aiming at structuring the text of relevant regulations in the electric power industry, this study proposes a named entity recognition model based on the features of characters and binary phrases. In this model, the word embedding representation is obtained by using the BERT pre-trained language model fused with multiple features, and the Transformer model and conditional random field that introduce the relative position coding are used as the encoding layer and the decoding layer, respectively. The model proposed in this study is applied in entity type recognition, and it can achieve effective recognition with the accuracy of as high as 92.64%.

Key words: named entity recognition; BERT model; Transformer model; conditional random field

随着智能电网的不断革新, 信息技术与电力系统逐渐融合, 电力企业在数字化转型的方向上有了极大的进展, 在这一过程中产生了大量的行业相关数据, 也可以称为是数字资产, 数字资产是企业或机构在生

① 基金项目: 国家重点研发计划 (2021YFE0102400)
收稿时间: 2021-09-07; 修改时间: 2021-10-11; 采用时间: 2021-10-15; csa 在线出版时间: 2022-05-26

产、运营、管理过程中累积的对企业或机构具有利用价值的数字化信息和内容, 通过对数字资产的组织加工, 可以优化企业内容管理架构, 促进企业运营模式改革, 从而提高企业收益. 简单地将海量的数字资产存储在各种存储介质中并且不采取任何管理措施, 企业的数字资产无法体现其自身的任何价值, 为了发挥企业数字资产的最大价值, 数字资产管理应运而生. 数字资产管理是对数字资产的创建、采集、组织、存储、利用和清除过程加以研究并提出的相应方法的统称.

将电力行业数据进行有效的组织管理可成为电力企业实现数字资产商品化[1] 的指导方法, 采用合适高效的分类管理方式不仅进一步加快电力企业数字化转型步伐, 也可以推动电力企业发掘新的利润增长点.

1 相关工作

知识图谱是谷歌公司于 2012 年首次提出的概念. 知识图谱的本质是一个与传统数据库不同的大型语义知识库, 知识库中主要涉及到的内容为数据中的实体与关系, 一个构建好的知识图谱可以用来辅助进行问答, 数据分析和决策等应用[2]. 知识图谱的构建包含知识抽取, 知识融合等内容, 其中知识抽取是构建知识图谱的核心环节. 将知识图谱的构建方法应用到电力企业所产生的庞大数据, 可促进对数据的高效分类并且促进实现数字资产商品化, 提升数字资产的价值.

命名实体识别[3] 是构建知识图谱最重要的也是最为根基的一环, 该任务旨在从文本中抽取待分类的命名实体并标注其类型, 通用的命名实体任务一般包括人名、地名、和机构名等. 实现特定领域的命名实体识别需要将领域内特定类别的实体识别出来, 该任务最早使用人工编写规则的方式进行识别, 例如通过制定有限的规则和模式, 从文本中自动匹配这些规则或模式的字符串, 并标记为各类命名实体, 不过随着数据集的越来越复杂, 用人工制定的有限规则识别日益增长的命名实体是非常困难的. 因此基于统计机器学习的方法获得了越来越广泛的关注, 使用统计机器学习的方法大致可分为以下几个步骤: 选择适合文本序列的模型, 使用合适的文本特征来增强模型的特征捕获能力. 并且将命名实体任务转化成为序列标注任务, 也即对序列中的每个字符都有多种标签类别的可能与之对应, 模型所要做的就是为每个字符分配可能性最大的分类标签, 从而使实体被标注为正确实体类型标签, 完成命名实体任务.

近年来, 随着基于神经网络模型的深度学习方法成为了机器学习中十分热门的方向, 其中利用语言模型等任务所得到的预训练高维词向量来作为词语的表示方法更是加强了神经网络模型的表示能力, 这样的表示不仅缓解了独热向量的数据稀疏问题, 还使得稠密的向量具有一定的语义表示能力. 从 Word2Vec 开始, 寻找一个有效的词向量表示成为了自然语言处理重要的研究方向, 文献 [4] 使用 BERT 模型在大规模语料中进行自监督的预训练, 得到每个字关于上下文的表示, 并且通过多层堆叠的 Transformer 模型, 可以使得文本的向量表示具有动态的, 上下文相关的特点, 这样的特点可以缓解过去静态的词向量无法解决词语歧义的问题.

郭军成等人[5] 利用 BERT 嵌入 Bi-LSTM 实现了对简历数据的命名实体识别, 吴超等人[6] 利用了 Transformer 混合 GRU 在电力调度领域进行了命名实体识别, 谢腾等人[7] 将 BERT 嵌入到 Bi-LSTM-CRF 中, 在 MASR 通用数据集中获得了显著的效果, 赵丹丹等人[8] 利用了多头注意力机制和字词融合实现人民日报中的通用领域命名实体识别. 与此同时, 近年来由于 Transformer[9] 使用自注意力机制在预训练任务上和机器翻译任务上表现十分出色, 如何将 Transformer 较好的适配到其他任务也成为了一个热点的研究方向, 韩玉民等人[10] 利用 Transformer 实现材料领域的英文命名实体识别, 何孝霆等人[11] 利用 Transformer 来捕捉文本的特征从而判断文本的真实立场.

本文提出了一种融合字符和二元词组特征, 通过 BERT 预训练模型得到上下文语义特征, 然后嵌入改良位置编码表示的 Transformer 模型的命名实体识别方法, 较好地实现了电力领域命名实体识别任务.

2 嵌入 BERT 的 Transformer NER 架构

2.1 CB-BRTC 模型

本文提出的模型 CB-BRTC 如图 1 所示, 可分为 4 部分, 第 1 部分为特征表示, 本文提出了一种基于字符级别的混合二元词组作为特征的输入. 第 2 部分为 BERT 模型, 通过使用中文的 BERT 模型可以得到上下文语义的表达, 将字符混合二元词组的特征通过 BERT 模型得到具有上下文语义表示的词嵌入向量. 第 3 部分为改进 Transformer 模型的编码层, 使用多头

自注意力机制自动捕捉文本在不同语义空间的表达, 使用相对位置编码融入词向量中. 第 4 部分为解码层, 通过使用条件随机场解码序列的输入, 从而得到最终的标签序列.

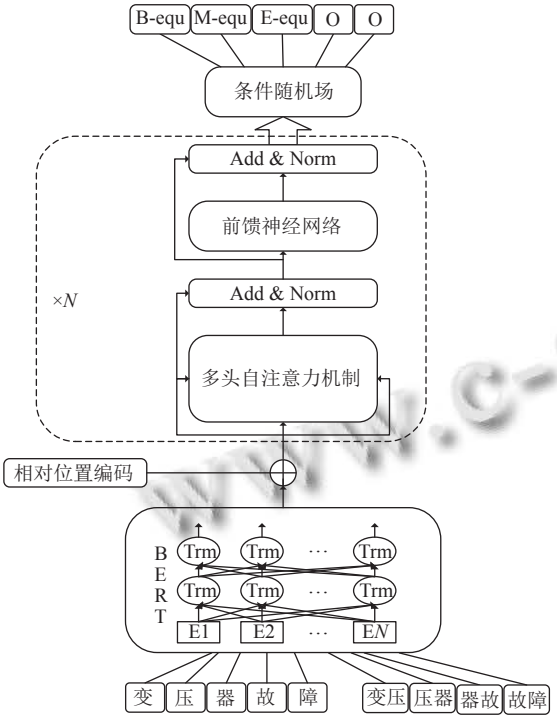

图 1 CB-BRTC 模型结构

算法的流程为:首先通过对收集来的电力政策数据进行预处理, 得到字符序列和字符的二元词序列, 然后将字符序列和二元词组序列通过使用 BERT 的词表映射, 使文本通过 BERT 计算每个字符的上下文语义表示的词向量, 随后使用相对位置编码, 融合词向量表示, 通过使用 Transformer 的多头自注意力机制捕捉序列在不同语义空间的表达, 再经过前馈神经网络进行融合, 最终得到关于命名实体识别任务的编码表示. 最后将该编码表示通过条件随机场解码得到符合序列标签转移规则的标签表达, 同时为了高效解码, 使用维比特算法得到分数最高的路径从而选择合适的标签序列.

本文使用的结构与前人主要的区别在于本文选择了对中文命名实体识别效果具有高效并且容易得到的字符融合二元词组特征, 利用 BERT 得到符合上下文语义的动态词向量使得句子的表达含义相较 Word2Vec 的静态词向量更加准确. 并且使用改进 Transformer 位置编码的编码结构, 通过使用相对位置编码来使得模型更易捕捉文本的前后关系.

2.2 BERT-WWM 模型

BERT 模型是一种基于自监督训练任务的预训练语言模型, 具有三大特点, 即使用海量的数据, 巨大的模型, 和使用强大的算力得到. BERT 本身采用多种 Transformer 堆叠而成, 使用了包括掩码语言模型和下一个句子预测的两个预训练任务.

BERT 使用词向量, 块向量和位置向量之和来表示输入. 通过使用掩码语言模型和下一个句子预测预训练任务来完成自监督训练. 掩码语言模型任务中, BERT 使用了 15% 的掩码比例, 将输入序列的 15% 的子词进行遮盖, 在这 15% 的遮盖子词中, 有 80% 的概率使用“[MASK]”标签来替换, 有 10% 的概率使用词表中的随机词来替换, 有 10% 的概率保持不变. 下一句预测是指利用文本中天然的句子顺序, 通过控制正负样本的比例在 1:1, 即使正确句子顺序和错误句子顺序的比例为 1:1, BERT 需要判断后一个句子是否是前一个句子的下一个句子, 从而学习到两段输入文本之间的关联.

BERT-WWM[12] 是在 BERT 的基础上使用进阶的预训练任务, 进一步提升预训练任务的难度, 从而使预训练模型具有更加有效的语义表达信息. BERT-WWM 使用的预训练任务是整词掩码的任务, 通过使用哈工大开发的 LTP 工具完成对语料的分词, 在进行整词掩码时, 将整个词语进行遮盖, 从而使得模型学习难度加大, 获得更好的语义表达形式.

2.3 编码结构

Transformer 是一种特殊的利用全连接的多头自注意力机制模型, 完整的 Transformer 是由编码结构和解码结构[9] 组成的, Transformer 的编码结构通过使用多头自注意力机制捕捉命名实体识别任务的文本在多个语义空间的表达和不同语义空间中文本序列中不同字符之间的关系. 本文使用的 Transformer 结构引入相对位置信息, 其编码单元的结构如图 2 所示.

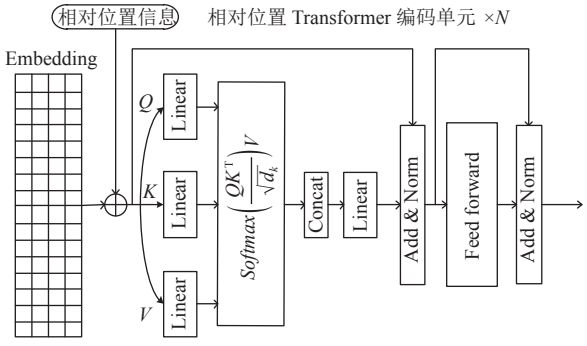

图 2 相对位置 Transformer 结构

在 Transformer 结构中使用的多头注意力机制如式 (1) 所示:

$$MultiHead(Q,K,V)=Concat(head_1,\cdots,head_n)\cdot W^o \quad (1)$$

其中, W^o是可以学习的参数, $Concat(head_1,\cdots,head_n)$指的是每一个$head_i$的拼接. 该式表示当文本序列经过多头自注意力的映射后, 可以通过一个矩阵的转换, 使多头注意力的输出进行降维. 多头自注意力中每个头$head_i$的表达式如式 (2) 所示:

$$head_i=Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \quad (2)$$

$$Q,K,V=HW_Q,HW_K,HW_V \quad (3)$$

其中, Q, K, V在原始 Transformer 结构中是由式 (3) 中的向量表示得到, H 为输入的序列矩阵, W_Q, W_K, W_V是可训练的参数, 通过多头自注意机制可以反映出来每个字符和其他字符的关系, 使用 Transformer 可以有效缓解卷积神经网络的专注于局部性的特点和循环神经网络梯度消失所导致无法实际捕捉长距离依赖关系的问题, 拥有比卷积神经网络和循环神经网络更好的特征捕捉能力.

Transformer 编码单元利用残差网络和层正则化来缓解深度神经网络中经常会遇到的退化问题, 具体的实现如式 (4) 所示:

$$Sublayer=LayerNorm(x+Sublayer(x)) \quad (4)$$

其中, x为通过复杂网络结构前的输出, 在 Transformer 中指多头自注意力映射前的输出或者通过前馈神经网络之前的输出, $Sublayer(x)$表示通过复杂结构之后的输出, 在 Transformer 中指多头自注意力映射拼接降维后的输出或者通过前馈神经网络的输出.

在每个多头注意力和残差连接后都会接一个位置全连接前馈神经网络, 表达式如式 (5) 所示:

$$FFN(x)=\max(0,\ xW_1+b_1)W_2+b_2 \quad (5)$$

其中, W_1和W_2为可学习的变换矩阵, b_1和b_2为可学习的偏置, 由于采用自注意机制来捕捉文本序列之间的关系, 其本身并没有可以感知位置的结构, 所以在 Transformer 的输入部分引入位置编码, 位置编码可以使用可训练的矩阵表达, 也可以通过事先设置好格式得到. 文献 [9] 采用了绝对位置编码, 如式 (5) 和式 (6).

$$PE(pos,2i)=\sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \quad (6)$$

$$PE(pos,2i+1)=\cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \quad (7)$$

其中, pos是位置, i是位置编码的第i维度, d是输入的维度, 原始的 Transformer 采用三角函数来将绝对位置进行编码. 但是有研究[13] 证明, 使用绝对位置编码会使得 Transformer 的位置感知能力丧失对方向的判断, 相对位置的编码结构可以使得模型对位置的感知更加敏感, 因此本文采用相对位置编码从而使得 Transformer 模型更适合命名实体识别这一任务. 不仅如此, 原版的 Transformer 模型存在一定的冗余参数, 因此本文将原版公式中涉及到两个可学习的参数的相乘结果用一个可学习参数替换, 使用相对位置编码和消除冗余参数后的多头自注意力机制计算公式如下所示:

$$Q,K,V=HW_q,H_{d_k},HW_v \quad (8)$$

$$\begin{aligned}R_{t-j}=\Bigg[&\sin\left(\frac{t-j}{10000^{\frac{2\times0}{d_k}}}\right),\cos\left(\frac{t-j}{10000^{\frac{2\times0}{d_k}}}\right),\cdots,\\&\sin\left(\frac{t-j}{10000^{\frac{2i}{d_k}}}\right),\cos\left(\frac{t-j}{10000^{\frac{2i}{d_k}}}\right),\cdots,\\&\sin\left(\frac{t-j}{10000^{\frac{2\times d_k/2}{d_k}}}\right),\cos\left(\frac{t-j}{10000^{\frac{2\times d_k/2}{d_k}}}\right)\Bigg]\end{aligned} \quad (9)$$

$$A_{t,j}^{rel}=Q_tK_j^T+Q_tR_{t-j}^T+uK_j^T+vR_{t-j}^T \quad (10)$$

$$Attention(Q,K,V)=Softmax(A^{rel})V \quad (11)$$

其中, t 是目标字符的位置, j 是文本中计算自注意力时每个字符的位置, Q_t和K_j是 t 位置的问向量和 j 位置的键向量, W_q和W_v是可学习的矩阵, H_{d_k}是由H以d_k为单位划分的每一个部分, 每一个部分使用一个自注意力的头来捕捉特征, u和v都是可学习的向量, R_{t-j}是相对位置编码的向量, 由正弦和余弦函数间隔填充, 从而可以捕捉前后字符出现的关系.

2.4 解码结构

为了充分利用不同标签之间的依赖关系, 本文采用条件随机场模型来捕捉序列标签之间的转移概率和发射概率, 从而得到更加符合标签顺序的标签序列, 一个序列的标签$y=l_1,l_2,\cdots,l_\tau$出现的概率由式 (12) 所示:

$$p(y|s)=\frac{\exp\left(\sum_i\left(W_{CRF}^{l_i}h_i+b_{CRF}^{l_{i-1},l_i}\right)\right)}{\sum_{y'}\exp\left(\sum_i\left(W_{CRF}^{l'_i}h_i+b_{CRF}^{l'_{i-1},l'_i}\right)\right)} \quad (12)$$

其中, h_i表示 Transformer 的输出, y'表示可能出现的所有任意标签序列, l_i表示标签序列中的第 i 个标签, $W_{\text{CRF}}^{l_i}$由l_i决定, 可以看作是发射概率的函数, $b_{\text{CRF}}^{(l_{i-1},l_i)}$由$l_{i-1}$和$l_i$决定, 可以看作是转移概率的函数. 由于计算所有可能形成的标签序列的得分计算量太大, 所以需要使用维特比算法来找得分最高的路径. 当给定一组训练数据$\{(s_i,y_i)\}|_{i=1}^N$, s 为输入的序列, y 为标签序列. 使用带有L_2正则化的极大似然损失来训练参数从而防止过拟合, 如式 (13) 所示:

$$L=\sum_{i=1}^{N}\log(p(y_i|s_i))+\frac{\lambda}{2}\Theta^2 \tag{13}$$

其中, λ表示L_2正则化的系数参数, Θ表示参数集合.

3 实验

3.1 实验环境和数据介绍

实验计算机的系统配置和主要程序版本如下: Linux 操作系统, Python 3.7, PyTorch 1.2, 16 GB 内存.

本文使用电力行业规章制度的标注文本实验, 对规章制度中的非结构化数据利用专家标注的方法构建标注数据, 并将数据分为训练集, 开发集和测试集, 数据集中的实体类型分为“机构单位”“电力设施”“政策原则”这 3 种实体类型. 采用“BMEO”的实体标注方式, 例如“机构单位”的实体用“B-AFF”“M-AFF”“E-AFF”来表示“机构单位”实体的开头中间和结尾. “O”代表该字符不属于 3 类实体的任一类. 数据集的规模如表 1 所示.

表 1 电力政策实体标注数据集

	训练集	开发集	测试集
句子条目数量	1011	335	336
最长句子字符数	274	271	226

3.2 命名实体识别效果对比

本文实验采用的评价指标为边界判断的准确率 (P)、召回率 (R) 和 $F1$ 值来表示模型对实体所在位置的定位, 用类型准确率 (type acc) 表示对实体判断的准确率. 公式的参数定义如下: T_P为模型识别正确的边界数量, F_P为模型识别出错误的边界的数量, F_N为该字符是相关实体的边界但模型没识别出实体边界的数量. 计算公式如下.

$$P=\frac{T_P}{T_P+F_P} \tag{14}$$

$$R=\frac{T_P}{T_P+F_N} \tag{15}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{16}$$

为了证明本文提出的模型的有效性, 本文使用了多种模型的对比实验来说明本文提出方法的有效性, 实验中, 我们分别对比了卷积神经网络中的膨胀卷积方法[14], 循环神经网络中的长短期记忆网络和双向长短期记忆网络[15], 原始 Transformer 模型与本文提出的 CB-BRTC 模型. 对比试验的结果如表 2 所示.

表 2 对比实验结果 (%)

模型	精确率	召回率	$F1$	Type acc
IDCNN-CRF	61.81	55.38	58.42	90.92
LSTM-CRF	62.79	61.02	61.89	91.69
BiLSTM-CRF	66.2	58.37	62.04	91.66
Transformer-CRF	61.19	59.93	60.55	90.88
CB-BRTC	64.14	66.67	65.38	92.64

从表 2 中可以看出, 本文提出的模型算法流程充分利用了输入的特征, 并具有高效的特征提取能力, 长短期记忆网络在一定程度上优于卷积神经网络, 因为其具有捕捉序列的长程依赖的能力, 原版的 Transformer 模型难以直接捕捉到序列的先后关系, 从而效果并不如长短期记忆网络. 通过引入相对位置编码, 本文提出的基于字符特征和二元词组特征的模型, 利用 BERT 构建符合上下语义的词嵌入表示, 并使用相对位置编码的 Transformer 模型得到序列中有方向感知的文本序列编码, 最后通过条件随机场解码, 取得了优于其他模型的效果, 在同样使用一层网络结构来对电力政策文本进行命名实体识别任务时, 本文提出的模型取得了最优的效果, 拥有比 IDCNN 方法高出 6.96% 的 $F1$ 值, 比长短期记忆网络方法高出 3.49% 的 $F1$ 值, 比原型 Transformer 高 4.83% 的 $F1$ 值.

本文利用字符和二元词组特征来作为神经网络的输入, 由于中文缺少像英文文本中的空格边界, 所以中文命名实体识别多采用字符特征来作为模型的输入, 而由于中文的语言习惯, 多为两字成词, 所以二元词组是实体中较为重要的特征, 本文融入这两种特征作为输入, 对命名实体识别的效果是有帮助的. 将字符特征和二元词组特征融合后通过 BERT 预训练神经网络, 可以将预训练神经网络中通过自监督训练得到的符合语义表示的字向量和词向量的高效表示迁移到本任务所使用的字符上. 利用增加了相对位置编码的 Transformer 神经网络捕捉针对命名实体识别在不同语义空间的特

征表达, 拥有比绝对位置编码的 Transformer 更好的方向感知. 最后使用机器学习中经典的条件随机场算法来捕捉标签的发射概率和转移概率, 从而减少差错, 得到更加合理的标签序列表达.

3.3 BERT 使用探究

目前现存的 BERT 使用方法有两种, 一是使用 BERT 当作特征, 固定参数, 不参与训练, 只训练 Transformer 和条件随机场的参数, 另一种是在 BERT 上做精调, 将 BERT 的参数也做训练. 本节主要探究了这两种不同方式之间的差异. 效果之间的不同如表 3 所示.

表 3 是否精调 BERT 模型结果 (%)

是否精调	精确率	召回率	*F*1	Type acc
是	64.14	66.67	65.38	92.64
否	62.65	60.30	61.45	92.11

在利用 Transformer 网络进行 BERT 的精调探究, 实验结果表示精调的结果要比不精调效果更好, 精调是为了在通用语义表示的基础上, 根据命名实体任务的特性进行领域适配, 使 BERT 模型与命名实体识别更加适配, 得到更加高效的电力行业规章制度文本表达, 这样的文本表达对识别命名实体有着更好的表达, 本文在探究 BERT 的使用方法中, 我们使用了更加细致的实验用来增强模型对电力命名实体识别的效果, 采用冻结和解冻的策略来反映 BERT 预训练模型对命名实体识别的促进作用, 冻结是指在训练的过程中, BERT 模型的参数不参与梯度下降算法进行迭代更新, 解冻之后, 随着模型一起更新, 结果如表 4 所示.

表 4 冻结解冻 BERT 参数 (%)

解冻epoch	精确率	召回率	*F*1	Type acc
0	64.14	66.67	65.38	92.64
5	65.69	63.48	64.57	92.46
10	67.15	64.61	63.44	92.46
20	67.36	60.67	63.84	92.49
100	64.75	63.30	64.01	92.45
200	65.86	56.37	60.75	92.18

实验结果表明使用 BERT 作为词嵌入层, 在一开始就解冻的策略是该模型进行命名实体识别任务最有效的策略. 经过使用不同超参数调试, 我们设置具有 BERT 学习参数为相对位置编码的 Transformer 学习率的 0.04 倍, 用来避免精调 BERT 所导致预训练模型的灾难性遗忘的现象[16].

在利用 BERT 进行模型的实验分析中, 本文对于电力行业规章制度的标注文本中的实体进行识别, 并探究了不同类型实体识别的效果, 分别采用精确率, 召回率和 *F*1 值 3 种评价指标. 精调的实验结果如表 5 所示.

表 5 不同实体类别识别结果 (%)

实体类型	精确率	召回率	*F*1
机构单位	79.53	78.92	79.23
政策原则	48.95	49.47	49.21
电力设施	53.85	67.47	59.89

实验结果表明模型对于机构单位的实体识别效果最好, *F*1 值为 79.23%, 机构单位的识别效果最好的原因可能是由于电力内部企业数量固定, 且政策中反复提及的电网机构单位模型更容易判断. 政策原则的识别效果最差, *F*1 值为 49.21%, 导致政策原则识别效果差的原因可能是因为实体类型过长, 并且政策原则种类繁多, 模型不能在有限的数据中得到很好的训练.

3.4 消融实验

为了探究本文所提出的方法中每个部分对电力政策文本进行命名实体识别的作用, 本节对 CB-BRTC 模型进行了消融实验, 分别通过去除二元词组特征, 去除 BERT 词嵌入层和去除 Transformer 中的相对位置编码信息的实验来探究每个模块对模型的作用, 实验结果如表 6 所示.

表 6 消融实验结果 (%)

模型	精确率	召回率	*F*1	Type acc
A1	60.88	67.04	63.81	91.99
A2	68.50	58.23	62.95	92.25
A3	58.83	69.31	63.64	92.95
CB-BRTC	64.14	66.67	65.38	92.64

表 6 中, A1 表示整个模型去除 BERT 后使用 Word2-Vec 在中文语料上的预训练词向量输入相对位置编码的 Transformer, 然后将编码结构的输出再通过条件随机场, A2 表示整个模型去除二元词组特征后通过 BERT 来构建词嵌入, 然后利用相对位置编码的 Transformer 对序列进行编码, 最后将编码的结果通过条件随机场, A3 表示模型去除 Transformer 中的相对位置编码, 替换为绝对位置编码, 通过使用 BERT 构建序列的词嵌入, 通过绝对位置编码的 Transformer 后将输出送入条件随机场. 表 6 的结果表明, 去除二元词组特征后的模型对电力政策文本的命名实体识别效果最差, *F*1 值为 62.95%, 并且当去除本文所提出模型中的任何一部分都会对模型的效果造成损伤. 这一实验结果说明本文

使用字符混合二元词组作为特征, 将 BERT作为词嵌入层, 利用相对位置编码的 Transformer 结构, 使用条件随机场作为解码输出, 对电力行业文本的命名实体识别效果有显著的效果.

4 结论与展望

本文提出了一种新颖的神经网络模型 CB-BRTC 模型来对电网企业的规章制度等文件进行信息提取, 识别出非结构化文本中的命名实体. 模型使用字符级别的向量和二元词组特征作为输入, 使用改进 Transformer 的结构作为编码器结构, 引入相对位置编码使得 Transformer 具有方向感知的功能, 最后使用条件随机场捕捉标签之间的转移概率和发射概率, 使得序列标注更合理. 本文提出的方法在电力行业规章制度上均比传统神经网络方法取得了更好的实体识别效果. 本文的方法可以促进将电力行业数据进行有效的组织管理, 通过进一步的结构化构建电力行业知识图谱, 不仅进一步加快电力企业数字化转型步伐, 也可以推动电力企业发掘新的利润增长点. 不过对于行业内的命名实体效果仍难以达到通用领域非常高的识别率, 找到更有效率的神经网络模型来促进行业数字资产管理, 这也是我们下一步工作的方向.

参考文献

1 余贻鑫, 栾文鹏. 智能电网述评. 中国电机工程学报, 2009, 29(34): 1–8. [doi: 10.3321/j.issn:0258-8013.2009.34.001]

2 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述. 计算机研究与发展, 2016, 53(3): 582–600. [doi: 10.7544/issn1000-1239.2016.20148228]

3 刘浏, 王东波. 命名实体识别研究综述. 情报学报, 2018, 37(3): 329–340. [doi: 10.3772/j.issn.1000-0135.2018.03.010]

4 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.

5 郭军成, 万刚, 胡欣杰, 等. 基于 BERT 的中文简历命名实体识别. 计算机应用, 2021, 41(S1): 15–19.

6 吴超, 王汉军. 基于 GRU 的电力调度领域命名实体识别方法. 计算机系统应用, 2020, 29(8): 185–191. [doi: 10.15888/j.cnki.csa.007595]

7 谢腾, 杨俊安, 刘辉. 基于 BERT-BiLSTM-CRF 模型的中文实体识别. 计算机系统应用, 2020, 29(7): 48–55. [doi: 10.15888/j.cnki.csa.007525]

8 赵丹丹, 黄德根, 孟佳娜, 等. 多头注意力与字词融合的中文命名实体识别. 计算机工程与应用. http://kns.cnki.net/kcms/detail/11.2127.TP.20210726.1521.024.html. [2021-08-01].

9 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

10 韩玉民, 郝晓燕. 基于子词嵌入和相对注意力的材料实体识别. 计算机应用. http://kns.cnki.net/kcms/detail/51.1307.TP.20210721.1659.010.html. (2021-07-22).

11 何孝霆, 董航, 杜义华. Transformer 及门控注意力模型在特定对象立场检测中的应用. 计算机系统应用, 2020, 29(11): 232–236. [doi: 10.15888/j.cnki.csa.007556]

12 Cui YM, Che WX, Liu T, *et al*. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021, 29: 3504–3514. [doi: 10.1109/TASLP.2021.3124365]

13 Yan H, Deng BC, Li XN, *et al*. TENER: Adapting transformer encoder for named entity recognition. arXiv: 1911.04474, 2019.

14 Strubell E, Verga P, Belanger D, *et al*. Fast and accurate entity recognition with iterated dilated convolutions. arXiv: 1702.02098, 2017.

15 Huang ZH, Wu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv: 1508.01991, 2015.

16 Sun C, Qiu XP, Xu YG, *et al*. How to fine-tune BERT for text classification? arXiv: 1905.05583, 2019.