高技术通讯 2022 年 第32 卷 第7 期:749-755

doi:10.3772/j.issn.1002-0470.2022.07.009

基于 BERT 模型的检验检测领域命名实体识别①

苏展鹏②* 李 洋* 张婷婷** 让 冉** 张龙波** 蔡红珍* 邢林林③**

(*山东理工大学农业工程与食品科学学院 淄博 255000)

(**山东理工大学计算机科学与技术学院 淄博 255000)

摘 要 针对检验检测领域存在的实体语料匮乏、实体嵌套严重、实体类型冗杂繁多等问题,提出了一种结合双向编码器表示法(BERT)预处理语言模型、双向门控循环单元(BIGRU)双向轻编码模型和随机条件场(CRF)的命名实体识别方法。BERT-BIGRU-CRF(BGC)模型首先利用 BERT 预处理模型结合上下文语义训练词向量;然后经过 BIGRU 层双向编码;最后在 CRF 层计算后输出最优结果。利用含有检测组织、检测项目、检测标准和检测仪器 4 种命名实体的检验检测领域数据集来训练模型,结果表明 BGC 模型的准确率、召回率和 *F*1 值都优于不加入 BERT 的对比模型。同时对比 BERT-BILSTM-CRF 模型,BGC 模型在训练时间上缩短了 6%。

关键词 命名实体识别;双向编码器表示法(BERT);检验检测领域;深度学习;双向门控循环单元(BIGRU)

0 引 言

命名实体识别作为自然语言处理中关键部分,是信息抽取的基础,其主要作用是在无序的文本中按不同的需求提取出特定有意义的命名实体。现广泛应用于智能回答、机器翻译、知识图谱等方面。同时命名实体识别技术是通用领域和特殊领域进行信息抽取加快信息化发展的基础。

目前命名实体识别在通用领域中已经得到了优良的效果[1],主要得益于通用领域的数据集的完善和模型方法的不断优化。起初主要是基于规则字典方法,这种方法耗时长、性能差、效果不佳,之后提出的基于统计的方法[2],像随机条件场模型(conditional random field,CRF)和隐马尔科夫模型(hidden Markov model,HMM),这类方法依靠通用领域精确的特征条件得到不错的训练结果[3-4]。但因为其需要人为参与操作多、成本高,逐渐被深度学习方法所取代。深度学习方法在降低人工成本的同时,也具有很强的领域适应性,可以在不同领域取得不错的训练效果。因此深度学习方法现已成为自然语言处理(natural language processing,NLP)领域的主流方法。Hammerton[5]将长短期记忆模型循环神经网络(long-short term memory,LSTM)模型应用在了命名实体识别任务中,取得了理想的实验结果。Huang 等人[6]提出了双向长短期记忆模型循环神经网络(bi-directional gate recurrent unit,BILSTM)搭配 CRF 模型,同时还融合了其他语言学特征以提升模型性能。王洁等人[7]将字向量作为输入,利用双向门控循环单元(bi-directional gate recurrent unit,BIGRU)搭配 CRF 模型提取会议名称的语料特征,发现与 LSTM 相比 GRU 的训练时间减少了 15%。Ma 和 Hovy[8]提出了一种基于 IDCNN-LSTM-CRF 的模型在不需要数据预处理的情况下可以提取字向量和字

① 国家重点研发计划(2018YFB1403302)资助项目。

② 男,1997 生,硕士生;研究方向:自然语言处理;E-mail:2389192403@qq.com。

③ 通信作者,E-mail:xinglinlin@sdut.edu.cn。

(收稿日期:2021-06-16)

符向量并在 conll 2003 数据集中得到了 91.21% 的准确率。Yue 和 Jie[9] 提出 lattice LSTM 模型将字符向量和字向量融合来解决一词多义问题,使模型更易识别实体边界。Radford 等人[10] 提出 OpenAI 生成式预训练(generative pre-training,GPT)模型,用 Transformer 编码代替 LSTM 来捕捉长距离信息,但是因为是单向的,无法提取上下文语义特征。Devlin 等人[11] 提出了能够更好获取字符、词语和句子级别关系特征的变压器双向编码器表示法(bidirectional encoder representation from transformers,BERT)预训练语言模型。通过 BERT 利用 Transformer 模型提升自身模型的抽取能力,能够更好地明确实体边界。

李妮等人[12] 提出 BERT-IDCNN-CRF 模型来解决 BERT 微调过程中训练时间过长的问题并在 MSRA 语料上 $F1$ 值能够达到 94.41%。顾亦然等人[13] 提出了 BERT-BILSTM-CRF 模型来解决专业电机领域中出现的小规模样本特征信息不足和潜在语义特征表达不充分的问题。

检验检测行业作为近几年出现的一个新兴服务业,到目前为止检验检测领域还没有相关的公开数据集,进行检验检测领域的命名实体识别训练缺乏训练数据。同时由于检验检测领域的特殊性,使得检测文本识别过程中存在以下难点:(1)检验检测领域实体搭配规则不明确,检测项目冗杂繁多;(2)检测组织在检验检测文本中存在大量缩写和简称;(3)检测仪器实体中存在含有功能、结构等多种名词叠加的长实体,这类实体难以准确识别实体边界。根据以上问题本文提出了 BERT-BIGRU-CRF 融合模型,在 BIGRU-CRF 轻模型的基础上引入了 BERT 预训练模型,通过 BERT 模型提前在大规模的中文语料库中进行训练,学习中文语义特征,以此来解决小规模数据集存在的语义特征不足的缺点,让模型可以在本文自行搭建的检验检测小数据集上取得更好的效果。同时 BERT 的 Transformer 模型可以提升模型的抽取信息能力,可以更好地识别检测实体边界,确定检测实体类型。而 BIGRU-CRF 模型可以在不降低性能的前提下减少模型训练时间。实验结果表明本方法可以较好地解决检验检测领域识别的难点,准确识别出检验检测领域实体,同时缩短训练时长。

1 BGC 检验检测领域实体识别模型

1.1 BGC 模型框架

模型主要是由 BERT 语言预训练层、BIGRU 编码层和 CRF 推理层 3 部分组成。首先将字符序列传入到 BERT 预处理模型中。BERT 会依据字符的位置向量、字向量和文本向量 3 部分加和得到最终的输入向量。经过编码获取到上下文语义信息,把融合后的向量输入给 BIGRU 网络模型。经过 BIGRU 的双向编码后,最终输出给 CRF 选择出最合理的标签序列输出。BGC 模型结构如图 1 所示。

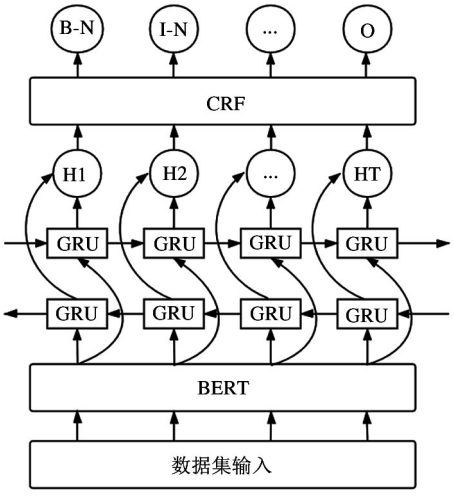

图 1 BGC 检验检测领域实体识别模型结构

1.2 BERT 预训练语言模型

BERT 作为一个预训练模型在处理特殊领域的命名实体识别任务时有着很好的语义表征效果,BERT 主要模型结构是由多组双向的 Transformer 编码构成。其无监督的方法,使得 BERT 可以通过预训练学习中文语义特点,通过迁移学习的方法转移到检验检测领域中来,以此来解决检验检测领域语料匮乏数据量少导致的文本特征信息不全面的问题。同时引用注意力机制对文本的上下文进行理解关联。因为其特殊的构造使得其可以使训练的结果更好反映出上下文语义关系,提取出检验检测领域特征信息,解决实体嵌套和冗杂的问题。同时双向

的 Transformer 结构要比传统 LSTM 模型可以更好地处理长期依赖问题。其作为编码器拥有强大的语言表征能力和特征提取能力[14],其结构如图 2 所示。

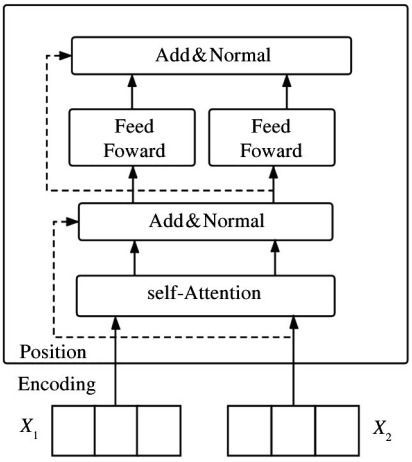

图 2　Transformer 编码器结构

本文利用 Transformer 本身引入的多头注意力机制,来多次计算学习不同的表示子空间,从而获取相关信息来提高模型在不同位置的专注度。经过 BERT 预处理模型对句向量处理后,通过多头注意力机制使句子向量在向量映射的不同情况下出现权重得分对比,使实体边界更易被识别。同时为了弥补自注意力机制不能抽取时序特征的问题,对词语中的位置进行编码:

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}}) \tag{1}$$

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{\text{model}}}) \tag{2}$$

其中,pos 表示词语在其句中位置;d 是 PE 的维度;d_k 是输入向量的维度;$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 都是自注意力机制(Self-Attention)计算出的字向量矩阵,$\boldsymbol{QK}$ 的乘积表示了此中某个词在整个句子中的关注程度。为了计算出向量在每一个映射的注意力向量:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = soft\max(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V} \tag{3}$$

将 Self-Attention 机制对之前的输出部分做一次残差连接和归一化处理将共享前馈神经网络在多个子空间中计算向量相似度:

$$X_{\text{output}} = X_{\text{embedding}} + Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \tag{4}$$

$$LayerNorm(x''_i) = \alpha(\frac{x''_i - \mu_L}{\sqrt{\sigma_L^2 + \varepsilon}} + \beta) \tag{5}$$

$$FFN = \max(0, xW_1 + b_1)W_2 + b_2 \tag{6}$$

最后为了使 BERT 在 NLP 任务中得到更好的表征效果,采用表征词语无监督方式的 Masked 语言模型和下一句预测 2 种任务方法[14]。Masked 语言模型采用随机屏蔽 15% 的信息,将被屏蔽的信息区分处理,把其中 80% 的信息替换成 Masked,10% 的被遮挡信息会被任意词替换,10% 的遮挡信息保持不变,以此来保证训练过程中模型可以更加准确获得词间信息。下一句预测则是采用二分类方法把任务中的句子分成上下连贯句子和非连贯句子 2 类。让模型预测判断两个句子间的关系,最后作出 IsNext 和 NotNext 2 种判断结果并进行标记,以此来获得句子间的上下关系。因此 BERT 模型可以在其他小数据领域中获得更好的全局表达效果,使得模型可以应用在检验检测这种语料匮乏的领域中。

1.3　BIGRU 网络层

GRU 和 LSTM 模型可以解决循环神经网络(recurrent neural network,RNN)模型无法长期记忆和反向传播中的梯度消失问题。但 GRU 作为 LSTM 的变体模型,其在保留了 LSTM 效果的基础上将遗忘门和输入门合并成了更新门,简化了模型参数。本文 LSTM 引入了 3 个函数来控制输入值记忆值和输出值,而 GRU 则是通过重置门和更新门 2 个门控制新的输入和记忆输入[15]。在 BERT 后搭配 BIGRU 模型,来增强模型训练速度和泛化性能,表达公式如下:

$$\begin{cases} z_t = \sigma(W_z[h_{t-1}, x_t]) \\ r_t = \sigma(W_r[h_{t-1}, x_t] \\ \tilde{h} = \tanh(W[r_t h_{t-1}, x_t] \\ h_t = (1 - z_t)h_{t-1} + z_t\tilde{h}_t \end{cases} \tag{7}$$

式中,z_t 为更新门 t 时刻的输出,r_t 代表重置门 t 时刻的输出,σ 为 sigmoid 函数,$\tilde{h}_t$ 和 h_t 分别表示 t 时刻的记忆内容和隐藏状态。GRU 结构如图 3 所示。

在单向的神经网络中,信息总是由前往后获取,只能采集到上游信息,为了让下游信息也能被获取,所以采用两条单向的 GRU 组成的正反两层的 BIGRU 模型,加强模型语义挖掘能力,能更好地理解前后文依存关系。

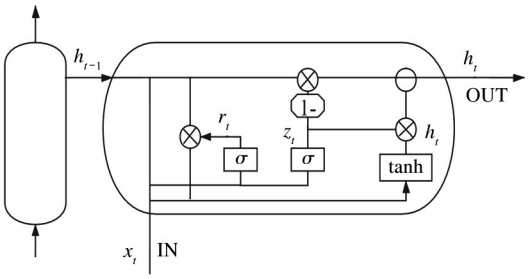

图 3 GRU 单元结构

1.4 CRF 层

虽然 BIGRU 模型的输出也能够评分出最优标签,但是只能在标签内判别。所以会出现很多错误标记和无用标记的问题。为此需要在 BIGRU 后加入一个 CRF 层来考虑标记信息间的依存关系。增加约束来确保预测标签的合理性,减少非法序列出现的概率。

对于最开始的输入字符序列 x 进行预测标签分布得到的标签序列 y;对于 BIGRU 层的输出矩阵 $\boldsymbol{p}$,其大小由一个句子中的单词数 n 和标记种类 k 所决定,其中 $P_{i,j}$、$A_{i,j}$分别为第 i 词在第 j 个标签的得分和转移矩阵中标签 i 到标签 j 的概率得分。最终序列 y 的得分如式(8)所示。

$$S(x, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_I, y_i \tag{8}$$

然后利用 Softmax 归一化处理得到最大的输出概率为

$$\boldsymbol{p}(y \mid X) = \frac{\exp(s(X, Y))}{\sum_{i=0}^{n} \exp(s(X, y))} \tag{9}$$

最后利用 Viterbi 算法[16]求得到所有序列上的预测中最优解作为检测实体结果,其公式为

$$y^* = \operatorname{argmax}(S(x, \tilde{y})) \tag{10}$$

2 实验分析

2.1 检验检测领域数据集

当前国内对检验检测领域的命名实体识别研究较少,缺少训练所需要的数据集。因此本实验通过网络爬虫的方法爬取了近 10 年来检验检测领域机构的检测类文本。整合检测类文本的特点和常用关键词,将命名实体分成以下 4 类,即检测项目、检测仪器、检测标准和检测组织。检验检测实体分类定义如表 1 所示。

表 1 检测实体识别类型及定义

符号标记	实体类别	实体定义	示例
M	检测组织	提供检测服务的公司、实验室、研究院	瑞士 ARL 公司、拜恩水质检测中心、英国牛津公司
P	检测项目	检测目标某一特征,检测产品某一性能,特色检测服务的名称。	硬度检测、热膨胀系数检测、爆炸极限测试、汞含量测定、熔体流动速率测定
S	检测标准	产品的国际标准、国标、行标、企标	GB/T 39665-2020 SN/T 4813-2017
E	检测仪器	检测所需要的设备和仪器	半导体 X 射线能谱仪、电感耦合等离子光谱仪、激光粒度仪、高速冷冻离心机、傅里叶变换红外光谱仪

根据其中 68 家网址的检验检测相关文本,总共搜集出 5.5 万条数据,为了搭建出本文命名实体识别所需要的数据集,经过了数据清洗和数据去重工作,得到有用数据条数 10 812 条。将得到的数据条数以 8∶1∶1 的比例分成训练集、测试集和验证集。同时本次实验采用的是 BIO 方法,命名实体以 B 为始端,I 为实体中间部分,O 为非实体部分。本文将实体分为了 4 类,因此出现的标注类别有 B-M、I-M、B-S、I-S、B-P、I-P、B-E、I-E、O 共 9 种。数据中各实体类别数量在数据集中的比例如表 2 所示。

表 2 不同实体在各数据集中的数量统计

实体种类	训练集	测试集	验证集	所有实体
检测组织	4152	406	458	5016
检测项目	4498	857	820	6175
检测标准	2296	342	336	2974
检测仪器	5770	861	881	7712

2.3 实验室环境和参数设置

实验室硬件设备为 CPU 采用 Intel(R) Xeon(R) Gold 6242R,GPU 采用 NVIDIA Tesla T4,内存为425 GB。实验室虚拟环境采用 Python3.6 的版本搭配 Tensorflow 1.14.0 的版本对模型进行训练和测试。本次实验采用 BERT-Base 模型,其含有 12 个 Transformer 层、768 个隐含层和 12 个多头注意力机制。为了使得下梯度下降的方向更加明确,减少训练时间。将批尺寸参数数值设置为 16,迭代轮次设置为 30。同时为了保证模型不会过拟合,将删除比例参数的数值设置为 0.5,学习率设置为 1×10^{-5}。

2.4 评价指标

本文采用准确率(precision,P)、召回率(recall,R)以及 *F*1(F1-sorce,F1)值作为模型性能的评价指标,具体公式为

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{13}$$

其中,*TP* 表示模型预测为正确且识别正确的样本个数,*FP* 表示模型预测为正确却未能识别的样本个数,*FN* 表示预测为错误的样本个数。

3 结果分析

为了验证在检验检测领域中本实验模型实体识别能力,在相同的实验环境下,本次实验加入了 3 个对照模型 BILSTM-CRF、BILSTM、HMM。通过搭建好的检验检测数据集进行模型训练,采用 *P*、*R*、*F*1 测试结果进行评估,每个模型在不同检测实体的实验结果如表 3 所示。

表 3 不同训练模型的对比结果(%)

实体类型	评价指标	BERT-BIGRU-CRF	BILSTM-CRF	BILSTM	HMM
检测项目	*P*	83.69	75.89	61.09	52.45
	R	78.92	71.83	65.58	57.03
	*F*1	81.24	73.80	63.25	54.64
检测标准	*P*	85.53	81.41	63.51	57.32
	R	85.80	70.19	75.32	58.97
	*F*1	85.67	75.39	68.91	58.14
检测组织	*P*	84.59	78.80	61.39	49.14
	R	78.20	65.59	65.00	38.78
	*F*1	81.27	71.59	63.14	43.35
检测仪器	*P*	88.23	81.98	64.08	57.05
	R	80.79	77.56	63.59	63.19
	*F*1	84.35	79.71	63.83	59.97

通过实验结果发现 BGC 模型是所有实验模型里面得分最高的,相较于其他模型得分有了明显提升。其中在 4 种实体中得分较高的实体是检测标准。其 *F*1 值达到了 85.67% 的分数,这类实体构成相对简单,在本次实验中取得了不错的识别得分。而其他类实体得分略逊色于检测标准实体。而 BILSTM-CRF 模型、BILSTM 模型、HMM 模型中 BILSTM-CRF 模型的 4 个实体 *F*1 得分更高,其识别效果仅次于 BGC 模型。

在加入 BERT 预训练模型后,BERT-BIGRU-CRF 和 BERT-BILSTM-CRF 2 种模型进行对比实验,观察实验结果和训练时长(min)如表 4 所示。

对比 BERT-BIGRU-CRF 模型和 BERT-BILSTM-CRF 模型,发现除了检测组织 *F*1 得分外,其余实体

表 4 BERT-BIGRU-CRF 和 BERT-BILSTM-CRF 在检验检测领域的 *F*1 值对比(%)和运算时间(min)对比

模型	检测组织	检测项目	检测标准	检测仪器	运行时间
检测指标	*F*1	*F*1	*F*1	*F*1	*T*
BERT-BIGRU-CRF	81.27	81.24	85.67	84.35	190
BERT-BILSTM-CRF	80.03	81.32	85.22	83.03	202

得分都略高于 BERT-BILSTM-CRF 模型。而 BGC 模型的训练时间缩短了 6%,证明模型简化后的性能依旧可以得到保证。

对其中表现较好的 BILSTM-CRF 模型和 BGC 模型的测试结果分析发现,BGC 模型的 *F*1 得分高的原因,在于其可以准确识别以下 3 种情况实体:(1)检测组织实体中存在的组织缩写如华测检测、岛泽测试等;(2)检测仪器实体中存在的长实体如电涡流式覆层厚度测量仪、溶体流动速率测定仪等;(3)检测项目中出现长实体如维卡软化温度检测、熔体流动速率测定等。而 BILSTM-CRF 模型在测试时,在出现第 1 种缩写问题时多次错误地将检测组织标注成检测项目。而在后 2 种情况出现时,BILSTM-CRF 模型标注结果中出现实体标注不全的问题。对比发现 BGC 模型在融合了 BERT 预训练模型和 BIGRU 轻模型后,通过分析学习上下文语义,结合学习的语料特征,使得模型可以结合上下文识别出检测组织的缩写,同时消除了歧义问题,明确实体边界使得检测项目和检测仪器实体中存在的一词多意和嵌套实体都被精准识别并区分出来。在本次小型检验检测数据集的训练情况下,BGC 模型得到了相对更好的训练结果。说明了本模型可以很好地解决检验检测领域中实体边界模糊和句式冗杂的问题。

同时发现,4 类实体中检测项目实体的 *F*1 得分相对较差,而检测项目实体又在检验检测领域中举足轻重。通过对测试结果整理分析后发现其主要原因是检测项目复杂多变。模型对检测项目中出现的新词的识别还有待加强。针对此实体特点后续可以再扩充检验检测数据集。在数据集内实体更加丰富的情况下,对检测项目类实体进一步定义区分。可以从检测范围、检测方法和检测产品角度去定义,更细致地划分检测项目实体,改善检测项目实体目前识别得分较低的情况。

4 结 论

本文从解决检验检测领域的命名实体识别问题出发,针对检验检测领域存在的语料匮乏、句式冗杂等问题提出了结合 BERT 预训练模型和 BIGRU-CRF 轻模型的检验检测领域命名实体识别方法 BGC。通过在 BGC 模型中加入 BERT 预训练模型后更好地识别出复杂实体的边界,对比未加入 BERT 的模型,该方法在 4 类实体中的 *P*、*R*、*F*1 结果有了显著的提升。而 BGC 模型使用 BIGRU 轻量模型后使得 BGC 模型比 BERT-BILSTM-CRF 模型的训练时间缩减了 6%,同时性能仍然优良。研究表明 BIGRU 模型确实可以在性能保持不变的情况下提高训练速度,减少训练成本。两个实验证明了 BGC 模型在检验检测领域的可行性。在后续数据集进一步补充、优化检验检测领域实体分类和定义后,将针对检验检测领域数据中更复杂的问题进行进一步的研究。

参考文献

[1] LI J, SUN A, HAN J, et al. A survey on deep learning for named entity recognition [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2020(99): 1-1

[2] 陈曙东,欧阳小叶. 命名实体识别技术综述[J]. 无线电通信技术,2020,46(3):251-260

[3] 张祝玉,任飞亮,朱靖波. 基于条件随机场的中文命名实体识别特征比较研究[C]//第四届全国信息检索与内容安全学术会议. 北京:中国中文信息学会信息检索与内容安全专业委员会,2008:102-107

[4] ZHOU G, JIAN S. Named entity recognition using an HMM-based chunk tagger[C]// Annual Meeting of the Association for Computational Linguistics, Philadelphia, USA, 2002: 473-480

[5] HAMMERTON J. Named entity recognition with long short-term memory[C]// Conference on Natural Language Learning at Hltnaacl Association for Computational Linguistics, Edmonton, Canada, 2003: 172-175

[6] HUANG Z, WEI X, KAI Y. Bidirectional LSTM-CRF models for sequence tagging[C]// Annual Conference of the North American Chapter of the Association for Computationa, Beijing, China, 2015: 13-16

[7] 王洁,张瑞东,吴晨生. 基于 GRU 的命名实体识别方法[J]. 计算机系统应用,2018,27(9):18-24

[8] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]// Annual meeting of the Association for Computational Linguistics 2016, Berlin, Germany, 2016: 1064-1074

[9] YUE Z, JIE Y. Chinese NER using lattice LSTM[C]// The 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018: 1554-1564

[10] RADFORD A, NARASIMHAK K, SALIMANS T, et al. Improving language understanding with unsupervised learning[EB/OL]. https://openai.com/blog/language-unsupervi-sed: OpenA, (2018-06-01), [2021-04-15]

[11] DEVLIN J, CHANG M W, LEE K, et al. Pretraining of deep bidirectional transformers for language understanding [C]// Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA, 2019: 278-286

[12] 李妮,关焕梅,杨飘,等. 基于 BERT-IDCNN-CRF 的中文命名实体识别方法[J]. 山东大学学报(理学版),2020,55(1):106-113

[13] 顾亦然,霍建霖,杨海根,等. 基于 BERT 的电机领域中文命名实体识别方法[J]. 计算机工程, 2021,47(8):78-83, 92

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems, Long Beach, USA, 2017: 6000-6010

[15] AYIFU M, WUSHOUER S, PALINDAN M, et al. Multilingual named entity recognition based on the Bi-GRU-CNN-CRF hybrid model[J]. *International Journal of Information and Communication Technology*, 2019, 15(3): 223

[16] STRUBELLE, VERGA P, BELANGER D, et al. Fast and accurate entity recognition with iterated dilatedconvolution [C]// Proceedings of the 2017 Cordially, Conference on Empirical Method in Natural Language Processing, Stroudsburg, USA, 2017: 2670-268

Named entity recognition in inspection and detection field based on BERT model

SU Zhanpeng*, LI Yang*, ZHANG Tingting**, RANG Ran**, ZHANG Longbo**, CAI Hongzhen*, XING Linlin**

(*School of Agricultural Engineering and Food Science, Shandong University of Technology, Zibo 255000)

(**School of Computer Science and Technology, Shandong University of Technology, Zibo 255000)

Abstract

Aiming at the problems of lack of entity corpus, serious nesting of entities, and multiple entity types in the field of inspection and detection, a named entity recognition method combining bidirectional encoder representation from transformers (BERT) preprocessing language model, bi-directional gate recurrent unit (BIGRU) bidirectional light coding model and random condition field (CRF) is proposed. The BERT-BIGRU-CRF(BGC) model first uses the BERT preprocessing model combined with contextual semantic training word vectors. Then it undergoes bidirectional encoding at the BIGRU layer. Finally it outputs the optimal result after calculation at the CRF layer. The model is trained by using the inspection and detection field data set containing four named entities of inspection organization, inspection items, inspection standards, and inspection instruments. The experimental results show that the accuracy, recall and *F*1 value of the BGC model are better than the comparison model without BERT. At the same time, compared with the BERT-BILSTM-CRF model, the BGC model shortens the training time by 6%.

Key words: named entity recognition, bidirectional encoder representation from transformers (BERT), inspection and detection field, deep learning, bi-directional gate recurrent unit (BIGRU)