



基于数控机床设备故障领域的命名实体识别

王欢 朱文球 吴岳忠 何频捷 万烂军

Named entity recognition based on equipment and fault field of CNC machine tools

WANG Huan, ZHU Wen-qiu, WU Yue-zhong, HE Pin-jie, WAN Lan-jun

引用本文:

王欢, 朱文球, 吴岳忠, 何频捷, 万烂军. 基于数控机床设备故障领域的命名实体识别[J]. 工程科学学报, 2020, 42(4): 476–482. doi: 10.13374/j.issn2095–9389.2019.09.17.002

WANG Huan, ZHU Wen-qiu, WU Yue-zhong, HE Pin-jie, WAN Lan-jun. Named entity recognition based on equipment and fault field of CNC machine tools[J]. *Chinese Journal of Engineering*, 2020, 42(4): 476–482. doi: 10.13374/j.issn2095–9389.2019.09.17.002

在线阅读 View online: <https://doi.org/10.13374/j.issn2095–9389.2019.09.17.002>

您可能感兴趣的其他文章

Articles you may be interested in

基于BiLSTM的公共安全事件触发词识别

Public security event trigger identification based on Bidirectional LSTM

工程科学学报. 2019, 41(9): 1201 <https://doi.org/10.13374/j.issn2095–9389.2019.09.012>

基于领域词典与CRF双层标注的中文电子病历实体识别

Clinical named entity recognition from Chinese electronic medical records using a double-layer annotation model combining a domain dictionary with CRF

工程科学学报. 2020, 42(4): 469 <https://doi.org/10.13374/j.issn2095–9389.2019.09.04.004>

基于周期势系统随机共振的轴承故障诊断

Bearing fault diagnosis by stochastic resonance method in periodical potential system

工程科学学报. 2018, 40(8): 989 <https://doi.org/10.13374/j.issn2095–9389.2018.08.013>

基于文本语料的涉恐事件实体属性抽取

Entity and attribute extraction of terrorism event based on text corpus

工程科学学报. 2020, 42(4): 500 <https://doi.org/10.13374/j.issn2095–9389.2019.09.13.003>

基于多域流形的行星齿轮箱局部故障识别

Localized fault identification of planetary gearboxes based on multiple-domain manifold

工程科学学报. 2017, 39(5): 769 <https://doi.org/10.13374/j.issn2095–9389.2017.05.016>

基于数控机床设备故障领域的命名实体识别

王 欢^{1,2)}, 朱文球^{1,2)}, 吴岳忠^{1,2)}✉, 何频捷^{1,2)}, 万烂军^{1,2)}

1) 湖南工业大学计算机学院, 株洲 412008 2) 湖南省智能信息感知及处理技术重点实验室, 株洲 412008

✉通信作者, E-mail: yuezhong.wu@163.com

摘 要 为了给数控机床故障的精准诊断提供保障, 延长数控机床使用周期, 以数控机床历史维修记录为研究对象, 对数控机床设备故障领域的命名实体识别进行了研究. 在分析历史维修记录中的故障描述特点后, 提出了一种基于双向长短期记忆网络 (Bidirectional long short-term memory, BLSTM) 与具有回路的条件随机场 (Conditional random field with loop, L-CRF) 相结合的命名实体识别方法. 首先, 对输入语句进行分词和标注, 使用 Word2vec 中的 Skip-gram 模型对标注语料进行预训练, 将其生成的字向量通过词嵌入层转化为字向量序列; 然后, 将字向量序列输入 BLSTM 学习长期依赖信息; 最后将句子表达输入 L-CRF 获取全局最优序列. 实验结果表明, 该方法明显优于其他命名实体识别方法, 为数控机床设备的智能检修与实时诊断任务打下了坚实的基础.

关键词 数控机床; 设备故障; 双向长短期记忆网络; 具有回路的条件随机场; 命名实体识别

分类号 TP391.1

Named entity recognition based on equipment and fault field of CNC machine tools

WANG Huan^{1,2)}, ZHU Wen-qiu^{1,2)}, WU Yue-zhong^{1,2)}✉, HE Pin-jie^{1,2)}, WAN Lan-jun^{1,2)}

1) School of Computer, Hunan University of Technology, Zhuzhou 412008, China

2) Hunan Key Laboratory of Intelligent Information Perception and Processing Technology, Zhuzhou 412008, China

✉ Corresponding author, E-mail: yuezhong.wu@163.com

ABSTRACT With the advent of intelligent manufacturing and big data, the Made in China 2025 Initiative and Industry 4.0 have been paying increasing attention to automation and intelligent industrial equipment. In the background of the present times, the complexity and intelligence of computer numerical control (CNC) machine tools have been continuously improved, and the types and descriptions of CNC machine tools' faults have increased, presenting serious challenges to equipment maintenance and diagnosis of CNC machine tools. In order to provide guarantee for accurate fault diagnosis of CNC machine tools, and to prolong the service life of CNC machine tools, it is necessary to improve the performance of named entity recognition system. Accordingly, the named entity recognition in the equipment and faults field of CNC machine tools were studied, taking the historical examinations and repair records of CNC machine tools as the research object. After analyzing the characteristics of fault description in the historical examinations and repair records, a named entity recognition method was proposed based on the combination of bidirectional long short-term memory (BLSTM) and conditional random field with loop (L-CRF). The first step is to input a sentence and segment and label the input sentence. The annotation corpus is combined with the pre-trained generated word vector by using Skip-gram model in Word2vec, and the word vector is converted into a word vector sequence through the word embedding layer. In the second step, the word vector sequence is integrated into the BLSTM layer to learn long term dependency information. The final step is to input the sentence expression into the L-CRF layer

收稿日期: 2019–09–17

基金项目: 国家重点研发计划资助项目 (2018YFB1700200); 国家自然科学基金青年科学基金资助项目 (61702177); 湖南省教育厅开放平台创新基金资助项目 (17K029); 智能信息感知及处理技术湖南省重点实验室开放课题资助项目 (2017KF07); 湖南省重点领域研发计划课题资助项目 (2019GK2133); 湖南省教育厅科学研究优秀青年资助项目 (19B147)

to obtain the global optimal sequence. The experimental results show that the method is superior to other named entity recognition methods, which lays a solid foundation for the intelligent maintenance and the real-time diagnostic tasks of CNC machine tools.

KEY WORDS computer numerical control machine tools; equipment failure; bidirectional long short-term memory; conditional random field with loop; named entity recognition

随着智能制造的快速发展和云计算^[1-2]时代的到来,“互联网+工业”成为了学者研究的重点,工业设备智能化对数控机床设备的检修与诊断提出了严峻挑战。数控机床设备智能检修与实时诊断是保障数控机床设备安全、稳定、经济运行的重要手段,是实现数控机床智能制造的关键。数控机床设备智能检修与实时诊断是通过整合海量、多源、异构的设备检修大数据,利用自然语言处理、语义匹配、知识图谱构建等技术对工业设备故障现象进行全面诊断,实现设备检修、诊断的智能化管理;数控机床设备故障命名实体的识别作为图谱识别、自然语言处理、语义匹配等复杂任务的基础工作,对数控机床设备智能检修与实时诊断起着至关重要的作用。为了提高命名实体识别系统的性能,为数控机床故障精准诊断提供保障,延长数控机床的使用周期,本文对数控机床设备故障领域的命名实体识别方法展开了研究,通过提高命名实体识别效果为后续数控机床智能检修和故障诊断工作打下了坚实的基础,进而防止意外故障带来的毁灭性事故和经济损失。

浅层机器学习和深度学习^[3]是通用领域的两种常用方法。浅层机器学习方法包括隐马尔可夫模型^[4]、最大熵模型^[5]和条件随机场^[6](Conditional random fields, CRF)等。俞鸿魁等^[7]在双层隐马尔可夫模型上进行实体识别,将不同的命名实体的识别结果融合到同一个理论模型中,提高了F值;何炎祥等^[8]提出CRF+特定规则模型,提高了召回率,改善了对实体识别的效果;王路路等^[9]针对维吾尔命名实体提出了半监督学习方法,减少了对人工特征提取的依赖。

近年来,命名实体的研究热点已从传统的机器学习方法转移到深度学习方法。与传统的机器学习方法相比,深度学习速度更快、泛化性更强,并且可以让计算机自主学习得到模式特征,将其融入实验模型从而减少对人工特征的依赖,因此使用深度学习进行命名实体识别的识别性能更好。Hochreiter与Schmidhuber^[10]提出了一种通过门限机制对历史信息进行过滤的LSTM,解决了循环网络中的梯度消失问题,但是LSTM只能获取下文信

息而无法获取上文信息。Graves与Schmidhuber^[11]构建了BLSTM模块,可以在输入的方向获得长时的上下文信息。杨红梅等^[12]提出了BLSTM-CRF命名实体识别模型,使标签结果更为合理。Lin等^[13]提出了多通道BILSTM-CRF模型在社交媒体中的新兴命名实体识别方法。Bharadwaj等^[14]在BILSTM-CRF模型上加入了音韵特征和attention机制,对有效的字符关注度更高;Li等^[15]针对包装专业领域语料匮乏的特点,提出了一个多层神经网络模型进行包装领域的命名实体识别,该模型可以自动学习分布式单词特征和部分语音特征,实现NER包装产品技术;易士翔等^[16]针对公共安全事件的触发词识别任务提出BLSTM与前向神经网络相结合的模型,在突发事件语料库上取得了较好的识别效果。

尽管通用领域深度学习的命名实体识别方法取得了较好的成果,但在数控机床领域中依然存在不足。陈秋瑗等^[17]提出了一种基于紧密度的命名实体识别,使用逻辑回归方法来计算相邻字串之间的紧密程度,提高了机械领域新词的准确率。在数控机床领域,学者对数控机床设备故障实体识别的研究较少,现有的研究主要针对数控机床特定部件存在的故障^[18-19];数控机床的历史维修记录是由工作人员撰写的描述数控机床异常的信息,该记录包括设备名称、故障描述、故障原因以及处理过程,这些都是数控机床的重要数据,自动抽取这些信息能够更加高效、精准的收集案例支持智能检修问答系统,而历史维修记录的重复利用主要受数据结构化程度的影响,因而对数据进行命名实体识别至关重要。

目前数控机床故障领域的实体识别存在以下几个难点^[20]:第一,没有基于数控机床领域的语料库;第二,没有数控机床故障领域的命名实体标注语料;第三,数控机床故障描述过于口语化,同一故障存在多种不同的描述。针对以上难点本文提出了一种基于双向长短期记忆网络与具有回路的条件随机场相结合的命名实体识别方法BLSTM-L-CRF,首先使用Word2vec^[21]对数据集进行预训练获取字向量,字向量通过词嵌入层得到字向量

序列, 然后将字向量序列输入 BLSTM 层, 经过 BLSTM 层处理捕获每个字对应的每个标签的分数, 最后将得到的标签分数输入 L-CRF 层, 并通过 L-CRF 层获得全局最优序列. 与 BLSTM-CRF 相比, 采用具有回路的 CRF, 约束性更好、识别效果更佳.

1 命名实体识别模型

1.1 BLSTM 模块

长短时记忆网络 (Long-short term memory, LSTM) 是一种改进后的循环网络. 与传统循环网络相比, LSTM 多了一个用来储存长距离信息的单元状态, 解决了梯度过长而产生的梯度弥散问题; LSTM 重复模块构造不同, 包含了四个交互层并以一种十分特别的形式进行交互; LSTM 中特殊设计的门结构使得模型能够决定丢弃信息、确定更新细胞、更新细胞状态. LSTM 的单元结构中内部参数可表达为:

$$\begin{aligned} i_t &= \sigma(W_i \cdot h_{t-1} + W_i \cdot x_t + b_i) \\ f_t &= \sigma(W_f \cdot h_{t-1} + W_f \cdot x_t + b_f) \\ o_t &= \sigma(W_o \cdot h_{t-1} + W_o \cdot x_t + b_o) \\ C_t &= f_t * C_{t-1} + i_t * \tanh(W_c \cdot h_{t-1} + W_c \cdot x_t + b_c) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \quad (1)$$

LSTM 有四个输入和一个输出, 输入处都经过激活函数, sigmoid 和 $\tanh$ 是 LSTM 的两个激励函数, 激活函数 σ 使用的是 sigmoid 函数, 值在 0 到 1 之间, 用以模拟门打开、关闭的效果. i_t 、 f_t 、 o_t 分别表示 t 时刻输入门、忘记门、输出门的输出, C_t 表示 t 时刻状态向量; W 为连接两层的权重矩阵, W_i 、 W_f 、 W_o 分别表示输入门、忘记门、输出门的权重矩阵, W_c 表示细胞状态下对于输入 x_t 的权重矩阵; b 为偏置向量, b_i 、 b_f 、 b_o 、 b_c 分别表示隐藏层的输入门、忘记门、输出门及细胞状态的偏置向量; x_t 为 t 时刻输入层的输入向量, h_t 、 h_{t-1} 分别为 t 时刻、 $t-1$ 时刻的输出, 其最终结果是一个高维实向量. LSTM 的实验包括三步, 首先, 通过忘记门决定丢弃信息; 然后, 通过输入门确定更新的信息; 最后, 采用 sigmoid 函数得到细胞状态的输出, 将输出结果与 $\tanh$ 相乘最终确定输出的信息.

LSTM 仅获取了文本的过去的信息, 但对数控机床设备故障命名实体识别任务而言, 设备故障描述的前后几个词对预测结果都有很大的影响, 获取上下文信息对数控机床设备故障的命名实体识别任务有很大的帮助. 为了获取上下文信息, 本文采用了双向 LSTM (BLSTM) 结构, BLSTM 由前

向的 LSTM 与后向的 LSTM 拼接而成. BLSTM 对每个句子分别采用正向和反向计算, 运用向量拼接将得到的两种结果进行向量拼接从而获得最后的隐层表示. BLSTM 结构如图 1 所示.

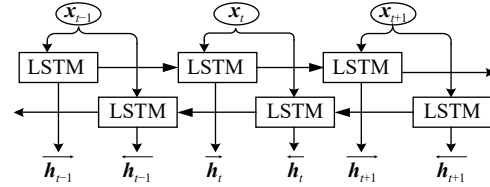


图 1 BLSTM 模型结构

Fig.1 BLSTM model structure

1.2 L-CRF 模块

条件随机场是 Lafferty 等^[22]提出的一种新的概率图模型, 该模型能从训练集数据中学习约束, 从而获得了全局最优序列, 同时解决了最大熵模型中存在的标注偏置问题.

线性链结构的 CRF 是最常见的 CRF, 它根据不同应用模式下 CRF 中的变量关系描绘出不同的图形. 例如, 可以选择树状条件随机场^[23]描绘具有阶层关系的变量. 针对数控机床历史维修记录中同一故障存在多种描述的特点, 本文采用 L-CRF^[24]架构来获取最优序列, 与传统的 CRF 相比, L-CRF 能够获取过去及未来时间点的信息, 使得到的序列准确率更高, L-CRF 架构如图 2 所示.

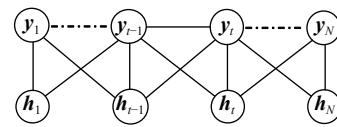


图 2 L-CRF 架构图

Fig.2 L-CRF architecture diagram

该架构图中, 输入值与输出值过去与未来的时间点包含在相互影响的关系中, 因而形成具体回路的图形架构. 为了能按照标准方法来推导模型, 按照构建联合树的步骤来检视图形, 根据图 2 建构了一个联合树, 通过对各时间点的联合树进行组合得到与 L-CRF 架构对应的联合树, 如图 3 所示, 图中圆节点代表团节点而方形节点代表分割节点.

将图形模型所代表的联合概率分布拆解成:

$$\begin{aligned} P_A(Y, H) &= \frac{\prod_{c \in C} \mu_c(h_c)}{\prod_{y \in Y} [\mu_y(h_y)]^{d(y)-1}} = \\ &= \prod_{t=1}^{N-1} \frac{\mu(h_t, y_t, y_{t+1}) \mu(h_{t+1}, y_t, y_{t+1})}{\mu(y_t, y_{t+1}) \mu(h_{t+1}, y_{t+1})} \end{aligned} \quad (2)$$

定义 H 为输入序列, Y 为一个联合树所有接口的



图3 L-CRF 架构所形成的联合树

Fig.3 Joint tree formed by L-CRF architecture

集合, N 为序列的长度; $d(y)$ 为与接口相邻的子图个数. 并且将 $\mu(h_t, y_t, y_{t+1})$ 、 $\mu(h_{t+1}, y_t, y_{t+1})$ 、 $\mu(y_t, y_{t+1})$ 和 $\mu(h_{t+1}, y_{t+1})$ 分别定义为:

$$\begin{aligned} \mu(h_t, y_t, y_{t+1}) &= \exp(\alpha_1^T \cdot g(h_t, y_t) + \beta_1^T \cdot g(h_t, y_{t+1}) + \\ &\gamma_1^T \cdot f(y_t, y_{t+1})) \mu(h_{t+1}, y_t, y_{t+1}) = \exp(\alpha_2^T \cdot g(h_{t+1}, y_t) + \\ &\beta_2^T \cdot g(h_{t+1}, y_{t+1}) + \gamma_2^T \cdot f(y_t, y_{t+1})) \mu(y_t, y_{t+1}) = \\ &\exp(\gamma_3^T \cdot f(y_t, y_{t+1})), \mu(h_{t+1}, y_{t+1}) = \exp(\alpha_3^T \cdot g(h_{t+1}, y_{t+1})) \end{aligned}$$

令 $\alpha_k^T = (\alpha_{k1}, \alpha_{k2}, \alpha_{k3}, \dots, \alpha_{kd})$, $\beta_k^T = (\beta_{k1}, \beta_{k2}, \beta_{k3}, \dots, \beta_{kd})$, $\gamma_k^T = (\gamma_{k1}, \gamma_{k2}, \gamma_{k3}, \dots, \gamma_{kd})$, 其中, α 、 β 为观察向量, γ 为特征函数, 将式(2)根据时间索引整理为:

$$P_A(Y, H) = \prod_{t=1}^{N-1} \psi(h_t, h_{t+1}, y_t, y_{t+1}) \quad (3)$$

当计算正规化项 Z 时, 可用以下的方法计算边界分布 $P(H)$:

$$P(H) = \sum_y P(H, Y) = \sum_y \prod_{t=1}^{N-1} \psi(h_t, h_{t+1}, y_t, y_{t+1}) \quad (4)$$

定义条件式对数相似度概率为 $P_A(Y|H, \theta) = \frac{1}{P(H)} \prod_{t=1}^{N-1} \psi(h_t, h_{t+1}, y_t, y_{t+1})$, 将上面所定义的相关项带入式中, 对式子整合后使用观察向量 a 取代原来的 α 和 β , 特征函数 b 取代原来 γ , 将式子改写成:

$$\begin{aligned} P(Y|H, \theta) &= \frac{1}{P(H)} \prod_{t=1}^{N-1} \exp(a^T \cdot f(y, y')) + \\ &\sum_{i=1}^A b_i^T \cdot g_i(h, y) = \frac{1}{P(H)} \exp\left(\sum_{k=1}^B \lambda_k^T \cdot F_k(H, Y)\right) \end{aligned} \quad (5)$$

其中 $\theta = \{\lambda_1^T, \lambda_2^T, \dots, \lambda_B^T\}$ 、 $a^T = (a_1, a_2, \dots, a_d)$ 、 $b_i^T = (b_{i1}, b_{i2}, \dots, b_{id})$ 、 $\lambda_k^T = (\lambda_{k1}, \lambda_{k2}, \dots, \lambda_{kd})$ 为所求模型参数, c 表示实体类别数, λ_k 表示特征函数对应的权值向量, F_k 表示特征函数, A 表示输入与输出向量间不同实体中具有关联性的个数, B 表示 L-CRF 中赋予定义的特征函数的总数.

1.3 BLSTM-L-CRF 模型

数控机床故障描述存在不同的表达方式, 例如: 发动机中的螺丝滑牙、发动机中的螺钉滑牙、发动机中的螺钉滑丝和发动机中的螺丝滑牙都是指同一故障, 在命名实体识别中易产生歧义, 为了解决上述挑战, 本文使用 BLSTM-L-CRF 模型进行命名实体识别. 在实验中, 将实体分为设备和实体两类, 用“Dev”表示设备, “Fau”表示故障. 首先根据 BIO 标注法对实体的定义进行标记, 即“B-Dev”

表示设备实体首字, “I-Dev”表示设备实体非首字部分, “B-Fau”表示故障实体首字, “I-Fau”表示故障实体非首字部分, “O”表示非实体部分, 因此定义标签集合为 (B-Fau、I-Fau、B-Dev、I-Dev、O). 然后使用 jieba 分词对输入语句进行分词, 分词后进行序列标注, 以“发动机中螺钉滑牙”为例进行序列标注, 其中发动机、螺钉是设备, 滑牙是故障现象, 标注后结果如表 1 所示.

表 1 句子序列标注方法

Sentence	Labeling	Sentence	Labeling	Sentence	Labeling
发	B-Dev	的	O	牙	I-Fau
动	I-Dev	螺	B-Dev		
机	I-Dev	钉	I-Dev		
中	O	滑	B-Fau		

获得标注语料后使用 Word2vec 进行预训练获取字向量, 在词嵌入层将其结果转化为字向量序列 X , 作为 BLSTM 层的输入. t 时刻序列正向输入 BLSTM 中得到的隐层输出为 $\vec{h}_t$, 逆向输入得到的隐层输出为 $\overleftarrow{h}_t$, 将两者拼接得到的 $h_t = [\vec{h}_t, \overleftarrow{h}_t]$ 包含上下文向量.

通过 BLSTM 输出的上下文特征信息 h_t , 得到有效的输出序列 $y = (y_0, y_1, y_2, \dots, y_n)$, 仅仅通过这样的方法进行分类还存在不足之处, 序列标注问题一般具有较强的依赖关系, 每个字的标签序列存在一定的局限性, 例如: 标签“B-Dev”作为设备实体的首字部分, 后面接的应该是“I-Dev”, 而不应该“I-Fau”. 只有 BLSTM 层是不够的, 针对此问题, 本文在 BLSTM 层加上处理标注序列的 L-CRF.

由于输入值与输出值之间存在过去与未来的时间点, 从而相互影响, 为此提出了具有回路的条件随机场(L-CRF), L-CRF 序列用来研究句子级别的序列特征, 将 BLSTM 层的输出结果输入 L-CRF 层, L-CRF 层对其结果添加约束, 从所有可能的标签序列空间中选出最佳序列路径, 获得全局最优序列. 本文的 BLSTM-L-CRF 模型架构如图 4 所示.

BLSTM-L-CRF 算法流程如下:

- 1: epoch $\leftarrow$ 1
- 2: while epoch < num_epoch do
- 3: t $\leftarrow$ 1
- 4: shuffle train and dev data

```
5:   while t < iteration do
6:       get a batch of train and dev data
7:       function forward
8:           forward pass of BLSTM-layer
9:           forward pass of L-CRF-layer
10:      end function
11:      function backward
12:          backward pass of BLSTM-layer
13:          backward pass of L-CRF-layer
14:      end function
15:  end while
16: end while
```

2 实验设计与结果分析

2.1 实验数据集

为了对 BLSTM-L-CRF 模型进行有效的评估, 本文以某工厂获取历年的数控机床历史故障维修记录作为数据集, 对数据进行了清洗和整理, 共获得 17485 条设备故障数据. 数据集按照 6 : 2 : 2 的比例进行划分, 即训练集 10491 条, 验证集 3497 条, 测试集 3497 条.

2.2 实验参数设置

Word2vec 是指将语言单词嵌入到向量空间从而得到词向量, 即该实验中的标注语料通过训练模型转换为向量形式. Skip-gram 和 CBOW 是 Word2vec 的两种常用模型, 文献 [25] 对两种模型进行对比实验得出以下结论: 当实验训练语料较少时, Skip-gram 模型效果更好, 当语料较多时则选用 CBOW 模型效果更好. 本文采用 Skip-gram 模型结合数控机床历史维修数据集预训练字向量, 通过对比实验获得最优参数, 如表 2 所示.

为了确定 BLSTM-L-CRF 模型的参数, 通过更

改源代码中的迭代次数进行对比实验发现, 迭代次数为 100 时识别效果最好. 通过采用不同的 Dropout 的值对模型进行了交叉验证, 结果表明 Dropout 值为 0.68 时达到最好的识别效果. BLSTM-L-CRF 模型参数如表 3 所示.

2.3 评价标准及结果分析

本文采用信息检索通用的评价方法作为评价标准, 即准确率(Precision)、召回率(Recall)和 F-测度值(F-measure, 指精确度和召回率的调和平均, 简称 F). 定义如下:

$$\begin{cases} P = \frac{n}{M} \times 100\% \\ R = \frac{n}{N} \times 100\% \\ F = \frac{2PR}{P+R} \times 100\% \end{cases} \quad (6)$$

其中, n 为正确识别的实体个数, M 为识别出的实体总数, N 为标准结果中的实体个数. 根据准确率、召回率和 F 值三个指标对模型的性能进行全面的评价.

对人民日报 1998 年 1 月份的新闻标注语料、MARA 语料和玻森 NLP 语料分别进行实验, 分析了该方法的可行性, 不同数据集的识别效果如表 4 所示.

从实验结果可知, 该模型在不同数据集上是可行的, 具有有效性.

使用不同模型在数控机床设备故障数据上进行对比实验, 不同模型下数控机床设备故障命名实体识别的识别效果如表 5 所示.

对比 L-CRF 模型与其他模型的实验结果发现, L-CRF 模型的准确率高, 但是单一的 L-CRF 模型的召回率和 F 值都比较低; 通过对比 BLSTM-L-CRF 模型与其他模型的实验结果, 发现该模型的

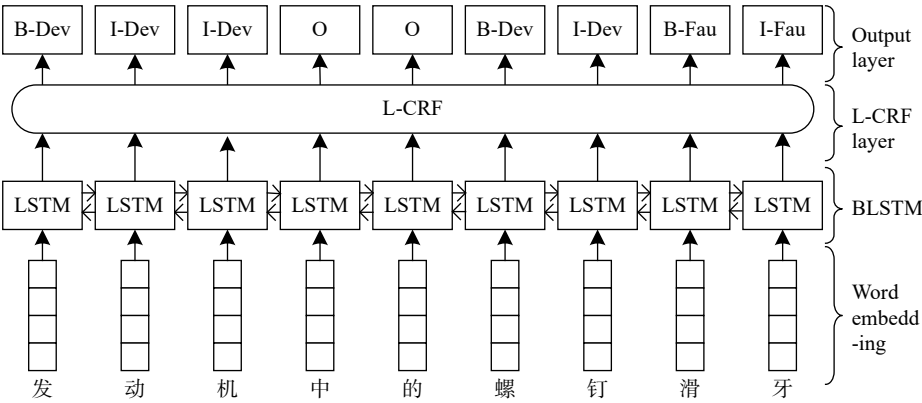


图 4 BLSTM-L-CRF 模型

Fig.4 BLSTM-L-CRF model

表2 Word2vec 的 Skip-gram 模型参数表

Table 2 Parameter list of Skip-gram model in Word2vec

Parameter	Value
Window size	10
Vector dimension	200
Minimum term frequency	5
Iterations	100

表3 BLSTM-L-CRF 模型参数表

Table 3 BLSTM-L-CRF model parameter table

Network layer	Parameter	Value
BLSTM	Learning rate	0.002
	BatchSize	20
	Iterations	100
	Dropout	0.68

表4 不同数据集在 BLSTM-L-CRF 模型中的识别结果

Table 4 Experiment result of BLSTM-L-CRF models in different data set

Date set	Precision/%	Recall/%	F-measure/%
People's daily corpus(1998)	83.07	83.40	83.23
MSRA corpus	82.23	80.35	81.28
Boson NLP corpus	79.45	80.18	79.81
CNC machine dataset	86.16	83.40	84.76

表5 BLSTM-L-CRF 与其他模型综合性能对比

Table 5 Comparison of performance of BLSTM-L-CRF and other models

Model	Precision/%	Recall/%	F-measure/%
CRF	85.45	69.87	76.88
L-CRF	85.92	72.54	79.16
LSTM	78.90	77.84	78.37
BLSTM	80.71	79.00	79.85
CNN-LSTM	83.62	80.07	81.81
BLSTM-CRF	81.54	80.41	80.97
BLSTM-L-CRF	86.16	83.40	84.76

准确率、召回率和 F 值都是最优的。在数控机床故障诊断领域,故障历史维修记录中的故障描述语言的上下文之间关联密切, BLSTM 能够获取上下文语义信息, L-CRF 包含过去和未来时间点的信息, 对一个序列而不是某个时刻的输出进行了优化, 使得标签结果顺序更合理, 都适用于数控机床领域, 由实验结果可知两种模型组合是本实验的最佳模型。

3 结论

(1)通过对 CRF 进行改进, 提出 L-CRF, 对上下文之间的关联进行更精准地推断, 通过对比实验发现, L-CRF 得到的序列更合理, 命名实体识别效果更好。

(2)通过使用 BLSTM-L-CRF 模型对不同语料进行实验, 发现该模型在不同数据集上都是可行的; 使用不同模型对数控机床数据集进行命名实体识别工作, 发现文章所提出的 BLSTM-L-CRF 模型识别性能最好。为了防止过拟合, 在模型中加入 dropout, 通过调整 dropout 值对实验模型进行对比实验, 发现 dropout 为 0.68 时, 防止过拟合的效果最好。

(3)提出采用 BLSTM-L-CRF 模型来提高命名实体识别的准确率, 在 BLSTM 层获取当前词语的上下文信息, 在 L-CRF 层对输入信息进行解码, 利用过去和未来时间点的信息获取最优标注序列, 该模型在数控机床设备及故障的命名实体识别上取得了较好的结果。后续将进一步完善语料库, 并对数控机床设备及故障进行实体关系抽取。

参 考 文 献

- [1] Zeng J D, Wang T, Jia W J, et al. A survey on sensor-cloud. *J Comput Res Dev*, 2017, 54(5): 925
(曾建电, 王田, 贾维嘉, 等. 传感云研究综述. *计算机研究与发展*, 2017, 54(5): 925)
- [2] Wang T, Shen X W, Luo H, et al. Research progress of trusted sensor-cloud based on fog computing. *J Commun*, 2019, 40(3): 170
(王田, 沈雪微, 罗皓, 等. 基于雾计算的可信传感云研究进展. *通信学报*, 2019, 40(3): 170)
- [3] Li J Y, Zhao Y K, Xue Z E, et al. A survey of model compression for deep neural networks. *Chin J Eng*, 2019, 41(10): 1229
(李江昀, 赵义凯, 薛卓尔, 等. 深度神经网络模型压缩综述. *工程科学学报*, 2019, 41(10): 1229)
- [4] Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[J/OL]. *arXiv preprint* (1998-03-27)[2019-09-17]. <https://arxiv.org/abs/cmp-lg/9803003>
- [5] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing. *Computat Linguist*, 1996, 22(1): 39
- [6] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons // *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*. Edmonton, 2003: 188
- [7] Yu H K, Zhang H P, Liu Q, et al. Chinese named entity

- recognition based on cascading hidden Markov model. *J Commun*, 2006, 27(2): 87
(俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别. *通信学报*, 2006, 27(2): 87)
- [8] He Y X, Luo C W, Hu B Y. Geographic entity recognition method based on CRF model and rules combination. *Comput Appl Softw*, 2015, 32(1): 179
(何炎祥, 罗楚威, 胡彬尧. 基于CRF和规则相结合的地理命名实体识别方法. *计算机应用与软件*, 2015, 32(1): 179)
- [9] Wang L L, Aishan W, Maihemuti M, et al. A semi-supervised approach to Uyghur named entity recognition based on CRF. *J Chin Inf Process*, 2018, 32(11): 16
(王路路, 艾山·吾买尔, 买买提·买买提, 等. 基于CRF和半监督学习的维吾尔文命名实体识别. *中文信息学报*, 2018, 32(11): 16)
- [10] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*, 1997, 9(8): 1735
- [11] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw*, 2005, 18(5-6): 602
- [12] Yang H M, Li L, Yang R D, et al. Named entity recognition based on bidirectional long short-term memory combined with case report form. *Chin J Tissue Eng Res*, 2018, 22(20): 3237
(杨红梅, 李琳, 杨日东, 等. 基于双向LSTM神经网络电子病历命名实体的识别模型. *中国组织工程研究*, 2018, 22(20): 3237)
- [13] Lin B Y, Xu F, Luo Z Y, et al. Multi-channel BiLSTM-CRF model for emerging named entity recognition in social media // *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Copenhagen, 2017: 160
- [14] Bharadwaj A, Mortensen D, Dyer C, et al. Phonologically aware neural model for named entity recognition in low resource transfer settings // *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, 2016: 1462
- [15] Li C Y, Wu Y Z, Hu F H, et al. Packaging domain-based named entity recognition with multi-layer neural networks. *NeuroQuantology*, 2018, 16(6): 564
- [16] Yi S X, Yin H P, Zheng H Y. Public security event trigger identification based on Bidirectional LSTM. *Chin J Eng*, 2019, 41(9): 1201
(易士翔, 尹宏鹏, 郑恒毅. 基于BiLSTM的公共安全事件触发词识别. *工程科学学报*, 2019, 41(9): 1201)
- [17] Chen Q Y, Cheng G, Li D, et al. Named entity recognition for mechanical design and manufacturing area. *Comput Eng Appl*, 2017, 53(20): 100
(陈晓媛, 程光, 李迪, 等. 机械设计领域的命名实体识别研究. *计算机工程与应用*, 2017, 53(20): 100)
- [18] Zhao X N, Feng Z P. Fault diagnosis of rolling element bearing based on ensemble empirical mode decomposition and cross energy operator. *Chin J Eng*, 2015, 37(S1): 65
(赵晓宁, 冯志鹏. 基于集合经验模式分解和交叉能量算子的滚动轴承故障诊断. *工程科学学报*, 2015, 37(S1): 65)
- [19] Zhang D, Feng Z P. Fault diagnosis of rolling bearings based on variational mode decomposition and calculus enhanced energy operator. *Chin J Eng*, 2016, 38(9): 1327
(张东, 冯志鹏. 基于变分模式分解和微积分增强能量算子的滚动轴承故障诊断. *工程科学学报*, 2016, 38(9): 1327)
- [20] Zhao W H, Zhang X, Lv D, et al. Technical status and strategies for domestic CNC machine tools. *Aeron Manuf Technol*, 2016, 59(9): 16
(赵万华, 张星, 吕盾, 等. 国产数控机床的技术现状与对策. *航空制造技术*, 2016, 59(9): 16)
- [21] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality // *Advances in Neural Information Processing Systems* 26(NIPS 2013). Lake Tahoe, 2013: 3111
- [22] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data // *Proceedings of the 18th International Conference on Machine Learning* 2001. Williamstown, 2001: 282
- [23] Yuan S, Tang J, Gu X T. A summary of scholars' portrait techniques in the open internet. *J Comput Res Dev*, 2018, 55(9): 1903
(袁莎, 唐杰, 顾晓韬. 开放互联网中的学者画像技术综述. *计算机研究与发展*, 2018, 55(9): 1903)
- [24] Zhu W Q, Liu Q. Conditional random fields with loop and its inference algorithm. *Comput Eng Appl*, 2008, 44(28): 180
(朱文球, 刘强. 基于条件随机场的上下文人类动作识别. *计算机工程与应用*, 2008, 44(28): 180)
- [25] Lai S W. Word and document embeddings based on neural network approaches[J/OL]. *arXiv preprint* (2016-11-18)[2019-09-17]. <https://arxiv.org/abs/1611.05962>