

面向机加工工艺规程文本的实体识别模型

董含笑, 李豫虎, 乔立红, 黄志成*

(北京航空航天大学机械工程及自动化学院 北京 100191)
(zc_huang@buaa.edu.cn)

摘要: 为实现非结构化工艺规程文本中关键信息的高效识别, 建立一种基于机加工领域词典和神经网络的命名实体识别模型. 首先, 结合机加工领域词典与 jieba 分词技术进行数据集的自动标注, 并在对工艺参数信息进行标注的过程中将数字和标志字母划分为一个分词单位以增强后续特征提取效果; 其次, 在 word2vec 词嵌入的基础上, 采用双向长短时记忆网络对文本进行特征提取; 最后, 采用条件随机场综合上下文逻辑以提高关键工艺信息的识别准确率. 在包含 431 条工步内容的数据集上, 对所提模型的识别效果进行实验, 结果表明, 所提模型的准确率、召回率和 F_1 值分别为 90.20%, 93.88% 和 92.00%, 在与领域内传统模型的对比上具有一定优势, 并使用 3 个不同工艺规程数据集验证了该模型的鲁棒性.

关键词: 双向长短时记忆网络; 条件随机场; 命名实体识别; 知识抽取

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.19732

Named Entity Recognition Method for Process Planning Text

Dong Hanxiao, Li Yuhu, Qiao Lihong, and Huang Zhicheng*

(School of Mechanical Engineering & Automation, Beihang University, Beijing 100191)

Abstract: To realize the efficient recognition of critical information in unstructured process planning text, a named entity recognition model based on technology dictionary and neural network is established. Firstly, the technology dictionary and jieba word segmentation technology are comprehensively combined to realize automatic annotation of datasets, especially, the number and its identification letters are recognized as one unit in the automatic annotation of process parameter data, which enhances the effect of subsequent feature extraction. Secondly, the bidirectional long short term memory network is used to extract the feature of text information based on word2vec. Finally, conditional random field model is used to synthesize contextual logic to improve the recognition accuracy of critical process information. To verify the effectiveness of the proposed model, 431 work steps are utilized as training sample. Experimental results show that the values of accuracy rate, recall and F_1 are 90.20%, 93.88% and 92.00% respectively, which has certain advantages compared with traditional models in the field. In addition, three experimental datasets from different technology books are tested, the results also show high robustness of the proposed model.

Key words: bidirectional long short term memory network; conditional random field; named entity recognition; knowledge extraction

收稿日期: 2023-05-31; 修回日期: 2022-09-13. 基金项目: 国家重点研发计划. 董含笑(2001—), 女, 硕士研究生, 主要研究方向为数字化制造、信息管理系统; 李豫虎(1998—), 男, 硕士研究生, 主要研究方向为数字化制造、信息管理系统; 乔立红(1961—), 女, 博士, 教授, 博士生导师, 主要研究方向为产品全生命周期管理、制造过程规划与管理、信息建模与集成; 黄志成(1991—), 男, 博士, 助理教授, 硕士生导师, 论文通信作者, 主要研究方向为增材制造、产品几何技术规范、知识工程、数字化与智能制造.

随着制造业的发展,企业已经积累了大量经过实践验证的工艺规程文件.其中蕴含着海量基础类、模板类和规则类的工艺知识,为工艺设计提供了指导.但这些工艺规程文件为非结构化的形式,通常依赖人工判别的方式进行存储和检索,效率低下^[1].

命名实体识别技术是解决上述问题的有效手段^[2].命名实体指的是工艺规程文件具有关键意义的信息,如工艺规程中的加工特征、粗糙度和加工方法等.随着计算机技术的发展,命名实体识别的研究重心逐渐由基于规则的方法向基于统计的方法和基于神经网络的方法转移.其中,基于规则的方法是早期命名实体识别研究中最有效的方式,其依赖人工制定的规则和权重,通过信息与规则的匹配进行实体识别,但其存在可移植性差、维护困难等问题.基于统计的方法是基于人工标注的语料集,通过序列标注进行命名实体识别,主流的方法包括条件随机场(conditional random field, CRF)^[3]、隐马尔可夫模型(hidden Markov model, HMM)^[4]等,但它们均对语料集的依赖性较强.基于神经网络的方法减少了对特征的依赖,同时提升了模型的泛化能力,在命名实体识别中得到广泛应用.常用的神经网络模型包含循环神经网络(recurrent neural network, RNN)^[5]、长短时记忆网络(long short term memory, LSTM)^[6]、卷积神经网络(convolutional neural network, CNN)^[7]等.

目前,学术界尝试将神经网络与基于规则和统计的方法进行结合,并取得了较好的效果.刘宇鹏等^[8]建立了一种基于双向 LSTM-CNN-CRF 的命名实体识别模型,应用于医疗领域和新闻领域中,得到了高于传统单一模型的结果. Liu 等^[9]提出一种融合神经网络与领域字典的中文分词方法,应用于临床医疗领域中,提升了识别的效率和准确性.陈义仁等^[10]提出一种基于神经网络和体素模板的骨骼受损类型自动判别方法,充分利用受损样本加强了神经网络模型的特征学习. Ji 等^[11]将药物字典与文本规则结合进行实体识别,进一步提升了模型性能,在电子病历命名实体识别方面取得了良好的识别效果.上述文献将神经网络与基于规则和统计的方法结合进行命名实体识别,并证明了其具有有效性和可扩展性.

机加工工艺规程文本具有专业性、经验性和多样性等特点,存在着大量专业术语(如加工方法中的车削、铣削和磨削),且包含大量工艺参数信息(如粗糙度和加工尺寸等),这些均为工艺设计的关键参考信息.因此,针对机加工工艺规程文本的特点,

本文构建了一种基于机加工领域词典和神经网络的命名实体识别模型,实现机加工工艺规程文本中关键信息的有效识别,为知识抽取任务奠定基础,以促进工艺知识的重用和共享^[12].

1 工艺规程文本特点分析

工艺规程文件一般包含工序过程卡片、工艺卡片和工序卡片等.本文针对工序卡片中以非结构化文本形式描述的工步内容进行命名实体识别研究.工步内容文本的特点如下.

(1) 句子长度较短.通常一条工步内容只包含一步操作,句子结构简单,以指导工人操作为主,属于短文本.

(2) 实体种类较少,专业性强.对于工步内容文本,一般仅包含加工方法、加工特征和粗糙度等关键工艺信息,用词范围较小,且属于专业术语.

(3) 参数信息重要.与其他领域不同,工步内容文本中包含的工艺参数信息在机加工工艺设计中意义重大,不可忽略,如粗糙度、加工尺寸和加工余量等.

综上所述,在进行数据预处理时,短文本中少且专的实体种类适合通过建立词典进行数据集的自动标注.此外,在自动标注时将数字和标示字母划分为一个分词单位有利于准确区分并提取工艺参数信息.

2 TD-BiLSTM-CRF 模型

针对工艺规程文本长度较短、术语专业性强和参数信息重要等特点,本文建立一种基于机加工领域词典和神经网络(technology dictionary-bidirectional LSTM-CRF, TD-BiLSTM-CRF)的命名实体识别模型.首先,参考工艺书籍和专家经验构建机加工工艺领域词典,词典中包含常见工艺术语,将词典与 jieba 分词技术相结合,实现数据集的自动标注.其次,使用 word2vec 将词语转化为向量,并构建 BiLSTM 模型捕获上下文语义特征,将向量序列转换为标注概率矩阵.最后,使用 CRF 模型生成标签转移矩阵,优化句子特征的提取效果,从而提升模型对工艺规程文本中关键信息的识别率. TD-BiLSTM-CRF 模型构建流程如图 1 所示.

2.1 数据集的自动标注

基于深度学习进行命名实体识别的方法对标注数据集的依赖性较强. jieba 分词技术可对数据

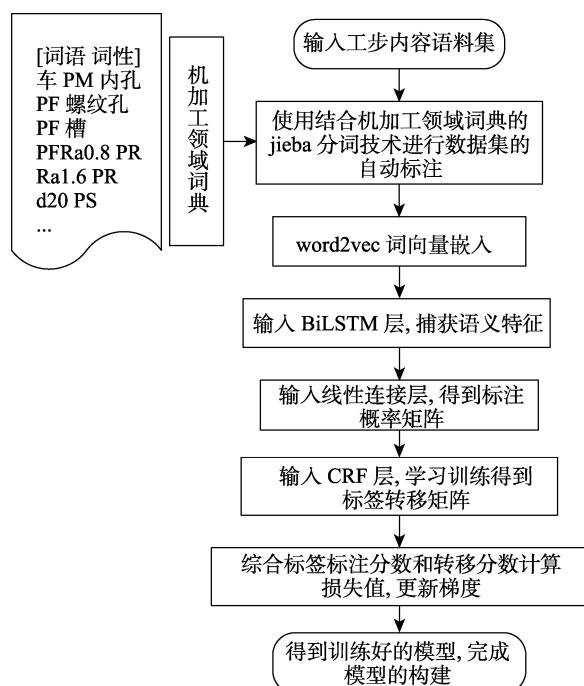


图1 TD-BiLSTM-CRF 模型构建流程

集进行粗略标注, 为进行深度学习、提升模型泛化能力提供数据支持. 针对机加工工艺规程文本实体种类较少、专业性强的特点, 本文采用结合机加工领域词典的 jieba 分词技术进行数据集的自动标注.

图2所示为数据集的自动标注过程. 首先, 本文构建了机加工工艺规程领域词典, 用于语料集中词语的切分和标注, 词典是分词算法的关键所在, 决定了最终分词的准确度. 其次, 基于统计词典构造一个 trie 树结构的前缀词典进行词条扫描, 并使用正则表达式, 以标点符号和非中文字符作为分界, 对输入的语料集按照子句进行切分, 得到所有切分可能, 根据切分的位置, 构造一个有向无环图 (directed acyclic graph, DAG). 最后, 根据动态规划算法在 DAG 上找到一条最大概率路径, 根据此路径得出最终切分形式, 按照词典进行标注. 对于未登录词(未在统计词典中出现的词), 加载 HMM 概率模型, 使用 Viterbi 算法找出最可能的隐状态序列, 进行词性标注.

2.2 工艺信息的识别

BiLSTM 模型虽然能较好地提取上下文语义特征, 但无法感知相邻标签间的依赖关系. 而 CRF 模型考虑标签之间的转移关系, 并计算得到全局最优标记序列, 因此, 本文结合 BiLSTM 模型和 CRF 模型, 用于提升机加工工艺规程领域命名实体的识别效果. BiLSTM-CRF 模型从输入到输出共有4层: Word Embedding 层, BiLSTM 层, 线性连接层

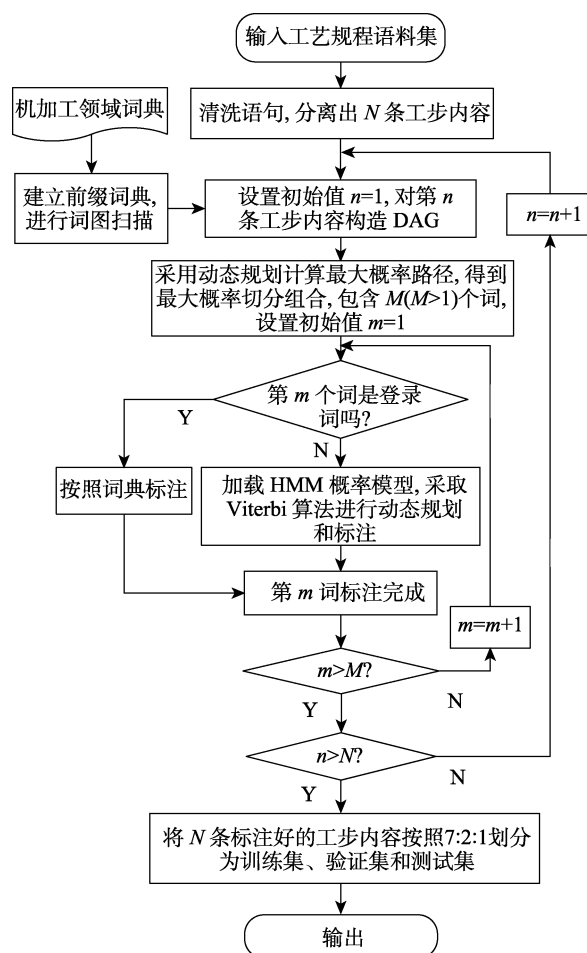


图2 数据集自动标注流程图

和 CRF 层. BiLSTM-CRF 模型结构如图3所示.

(1) Word Embedding 层

Word Embedding 是一个基于分布式假设的预训练模型. 本文采取 word2vec 方法进行词嵌入, 词嵌入过程如图4所示. 相较于传统的 one-hot 编码, word2vec 词嵌入一方面将字符转化为低维稠密向量, 节省了存储空间, 提升了计算能力; 另一方面, 捕捉全局上字词的相似度, 提升了模型泛化能力.

(2) BiLSTM 层

BiLSTM 由正和反2个 LSTM 组合而成, 其能够接收句子正向和反向的特征输入, 从而得到完整的上下文信息. 因此, 相较于单层的 LSTM, 双层的 LSTM 增强了识别过程中对语义特征的读取效果. BiLSTM 层的输入为降维之后的 n 维向量序列 $c = (c_1, c_2, \dots, c_n)$, 通过 LSTM 前后项特征提取之后, 分别得到前、后向隐状态序列 $h_L = (h_{L1}, h_{L2}, \dots, h_{Ln})$ 和 $h_B = (h_{B1}, h_{B2}, \dots, h_{Bn})$, 并在每个时间步 t 按位置拼接得到 $h_t = [h_{Lt}; h_{Bt}] \in \mathbb{R}^m$. 其中, m 表示拼接后的隐状态序列维度.

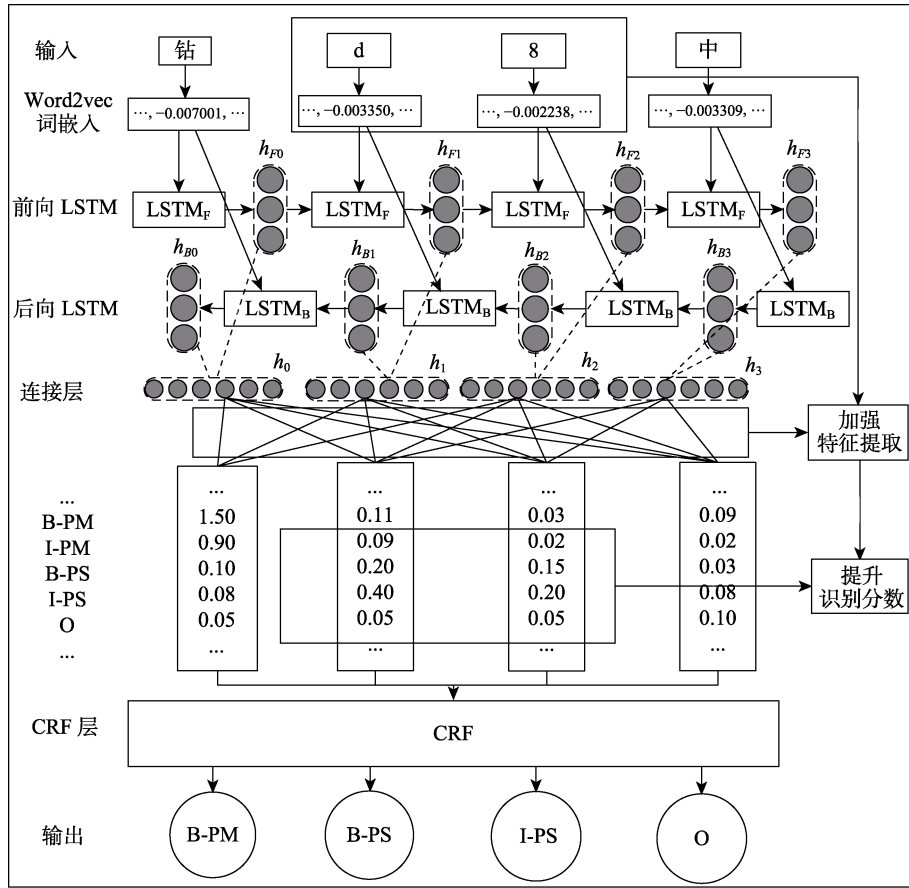


图 3 BiLSTM-CRF 模型结构

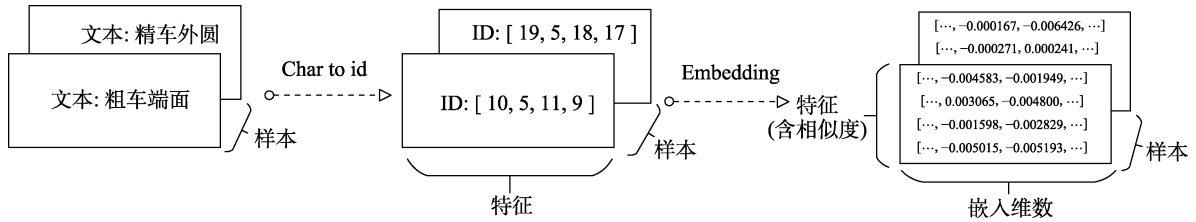


图 4 word2vec 词嵌入过程

(3) 线性连接层

经过 BiLSTM 层提取特征后，句子隐状态序列为 m 维，线性连接层将 m 维隐状态序列映射为 k 维， k 表示数据集的标签数量，该层的输出为每个字 x_i 分类到第 j 个标签的打分值 E_{ij} ，输出的标注概率矩阵 $E = (E_1, E_2, E_3, \dots, E_n) \in \mathbb{R}^k$ 。对于数字来说，由于在前面训练时，将数字和标志字母进行关联，无意义数字符号的特征学习转为了有意义数字符号的特征学习，进而提高了工艺参数的正确标签概率。

(4) CRF 层

CRF 层作用于句子级的序列标注，该层可以自动学习形成转移矩阵，即一个标签转移到另一

个标签的概率大小，使得在对工艺信息进行识别时，可以同时兼顾到每个字的标签打分信息和转移矩阵中标签之间的依赖关系。由于 CRF 层概率的约束，一个句子的开头可以为“B-PF”或“O-PF”，不能为“I-PF”。若一个句子的输入序列为 $x = (x_1, x_2, x_3, \dots, x_n)$ ，经过该模型的预测，输出的标注序列为 $y = (y_1, y_2, y_3, \dots, y_n)$ ，该序列每个字 x_i 标注为 y_i 的打分，即

$$S(x, y) = \sum_{i=1}^n E_{x_i, y_i} + \sum_{i=1}^n T_{y_{i-1}, y_{i+1}} \quad (1)$$

其中， E_{x_i, y_i} 表示经过 BiLSTM 层和线性连接层提取特征后输出的标签标注分数； $T_{y_{i-1}, y_{i+1}}$ 表示经过 CRF 层约束后输出的标签转移分数。进而可以利

用 Softmax 得到归一化后的概率为

$$P(y|x) = \frac{e^{S(x_i, y_i)}}{\sum_{i=1}^n e^{S(x_i, y_i)}} \quad (2)$$

其中, $S(x_i, y_i)$ 表示每条标注序列的打分;
 $P(y|x)$ 表示每条标注序列的概率, 选取其中概率最大的标注序列作为预测结果输出。

3 实例及结果分析

3.1 工艺规程数据集的获取与标注

(1) 数据获取

本文选用非结构化工步内容进行 TD-BiLSTM-CRF 模型识别效果的验证。工步内容数据集从相关教材、专著及手册中获取, 并加以扩展。本文数据集中包含 431 条工步内容, 共 6303 个字, 部分工步内容如图 5 所示。

粗镗 d80H9 内孔到 d77。
粗铣下端面, 保证尺寸 31 mm。
半精铣下端面, 保证尺寸 30.2 mm。
以毛坯底边为基准, 钻 d20 孔。
扩孔至 d21.8。
铰孔至 2 281, 保证孔的粗糙度。
钻 6-d9 mm 孔。
铰 6-d9 孔中对称轴上的两个孔至 d9.1H8 做定位孔。
以左端面及外圆固定, 粗车右端面及一半外圆尺寸 d70 mm。
粗车右端外圆使齿轮厚度 17.3 mm、d32.31 mm 并倒角 30° 调头, 固定。

图 5 部分工步内容

(2) 标签设计

为了对非结构化工步内容中的关键工艺信息进行准确识别, 本文设定了加工精度(processing accuracy, PA)、加工方法(processing method, PM)、加工特征(processing feature, PF)、加工尺寸(processing size, PS)和粗糙度(processing roughness, PR) 5 个实体标签。

(3) 样本标注

通过查阅资料, 综合专家经验, 本文构建了机加工工艺规程领域词典, 该词典涵盖了机加工工艺规程领域常见术语, 包括加工精度、加工方法和粗糙度等。

BIO 序列标注将每个实体标注为“B-X”“I-X”或者“O”。其中, X 表示上述定义的实体标签; B, I, O 分别表示该实体处于 X 的开头部分、中间部分及结尾部分和句子中的无关成分。本文采用结合机加工领域词典的 jieba 分词技术进行 BIO 自动标注, 标注结果示例如图 6 所示。

半 B-PA 精 I-PA 车 B-PM d B-PS 2 I-PS 2 I-PS m O m O 和 O d B-PS 2 I-PS 2 I-PS 8 I-PS m O m O 段 O 的 O 各 O 部 O 分 O 外 B-PF 圆 I-PF 。 O	半 B-PA 精 I-PA 镗 B-PM d B-PS 1 I-PS 4 I-PS 孔 B-PF 到 O d B-PS 1 I-PS 7 I-PS I-PS 8 I-PS , O R B-PR a I-PR 3 I-PR I-PR 2 I-PR 。 O	在 O d B-PS 3 I-PS 5 I-PS , O d B-PS 3 I-PS 7 I-PS , O d B-PS 4 I-PS 0 I-PS 外 B-PF 圆 I-PF 柱 O 面 B-PF 上 O 铁 O 键 O 槽 B-PF 。 O	调 O 头 O , O 精 B-PA 磨 B-PM 另 O 一 O 端 O 直 O 径 O 为 O d B-PS 3 I-PS 0 I-PS , O d B-PS 3 I-PS 7 I-PS 外 B-PF 圆 I-PF 柱 O 面 B-PF 。 O
---	---	--	--

图 6 BIO 标注结果示例

(4) 实体识别

将上述经过 BIO 序列标注后的数据集输入到后续的 BiLSTM-CRF 神经网络中进行训练。使用训练好的模型对未知工步内容进行预测, 部分识别结果如图 7 所示。实验结果表明, 该模型工步内容中包含的工艺信息识别完整, 且能准确识别工艺尺寸和粗糙度 2 种工艺参数。

{'entities': [{'end': 4, 'start': 3, 'type': 'PA', 'word': '粗'}, {'end': 5, 'start': 4, 'type': 'PM', 'word': '车'}, {'end': 10, 'start': 5, 'type': 'PS', 'word': 'd18.5'}, {'end': 14, 'start': 12, 'type': 'PF', 'word': '外圆'}, {'end': 20, 'start': 15, 'type': 'PS', 'word': 'd16.5'}, {'end': 39, 'start': 34, 'type': 'PR', 'word': 'Ra3.6'}], 'string': '调头, 粗车 d18.5 mm 外圆至 d16.5 mm, 长 31.5 mm, 粗糙度 Ra3.6'}
{'entities': [{'end': 14, 'start': 12, 'type': 'PA', 'word': '半精'}, {'end': 15, 'start': 14, 'type': 'PM', 'word': '镗'}, {'end': 18, 'start': 16, 'type': 'PF', 'word': '端面'}], 'string': '安装零件, 夹持 20 mm, 半精镗右端面, 留 0.3 mm 余量, 保证总长 98 mm'}
{'entities': [{'end': 4, 'start': 3, 'type': 'PM', 'word': '钻'}, {'end': 6, 'start': 4, 'type': 'PS', 'word': 'd8'}, {'end': 13, 'start': 10, 'type': 'PF', 'word': '中心孔'}], 'string': '倒角, 钻 d8 mm 定位中心孔, 作准尺寸'}
{'entities': [{'end': 9, 'start': 7, 'type': 'PA', 'word': '半精'}, {'end': 10, 'start': 9, 'type': 'PM', 'word': '车'}, {'end': 12, 'start': 10, 'type': 'PF', 'word': '外圆'}, {'end': 20, 'start': 13, 'type': 'PS', 'word': 'd66.946'}, {'end': 41, 'start': 36, 'type': 'PR', 'word': 'Ra1.8'}], 'string': '以两端面固定, 半精车外圆至 d66.946 并倒角 0.5×45°, 保证粗糙度 Ra1.8'}

图 7 标注结果示例

3.2 训练结果及对比分析

(1) 不同参数训练结果对比分析

模型中的参数配置会影响模型的性能和识别的准确率。本节探讨迭代次数、隐藏层维数和学习率等参数对模型的平均精准率、召回率和 F_1 值的影响。各个训练参数对模型的影响如表 1 所示。

图 8 所示为模型的精准率、召回率和 F_1 值随

表 1 各训练参数对模型性能的影响

参数名称	影响模型的性能
迭代次数	学习程度和优化速度
批次大小	泛化性能
隐藏层维数	精度
学习率	收敛速度

迭代次数的变化. 从图 8 中可以看出, 迭代约 10 次后, 模型快速收敛, 性能快速提升; 在迭代约 12 次后, 模型的性能不再大幅度提升, 趋于稳定. 因此, 在保证模型性能的前提下, 综合考虑时间成本和训练效果的稳定性, 本文选择迭代次数为 40.

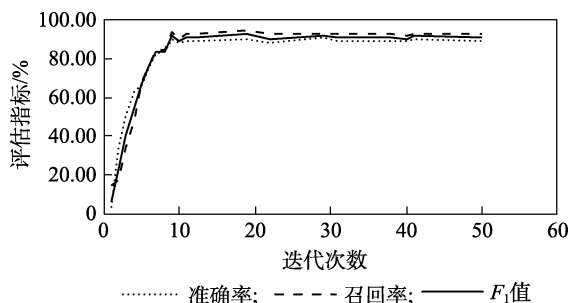


图 8 评估指标随迭代次数变化

图 9 所示为模型的 3 个评价指标随批次大小的变化曲线, 可以看出, 模型的评估指标值的变化趋势为先增大再减小再增加再减小, 当批次大小为 16 时, 达到近似最优, 因此, 选取批次大小为 16.

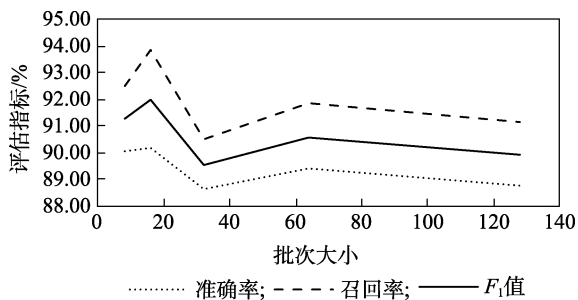


图 9 评估指标随批次大小变化

隐藏层的维数和学习率也会影响模型的性能, 表 2 为在迭代次数为 40 和批次大小为 16 的情况下, 改进后模型的 3 个评价指标随隐藏层维数和学习率的变化. 从表 2 中可以看出, 当隐藏层维数为 256、学习率为 0.003 时, 模型的性能最好.

因此, 在该数据集下, 取参数迭代次数为 40, 批次大小为 16, 隐藏层维数为 256, 学习率为 0.003, 此时, 可得到最优性能的 TD-BiLSTM-CRF 模型.

(2) 不同识别模型结果对比分析

将传统的 HMM 模型、CRF 模型和 BiLSTM

表 2 不同隐藏层维数和学习率下的评估指标 %

训练参数		准确率	召回率	F_1
隐藏层维数	学习率			
128	0.001	90.13	93.20	91.64
	0.003	88.67	90.48	89.56
	0.005	90.07	92.52	91.28
256	0.001	89.54	93.20	91.33
	0.003	90.20	93.88	92.00
	0.005	89.47	92.52	90.97

注: 粗体表示最优值.

模型与本文所提出的 TD-BiLSTM-CRF 模型进行对比分析. 4 种模型对 10 个不同标签的识别结果如表 3 所示.

分析 4 种模型的总体识别率: HMM 模型相比其他模型的效果较差; CRF 模型和 BiLSTM 模型识别的 F_1 值分别为 88.01% 和 89.98%, 对比准确率和召回率可以看出, 这 2 种模型可以识别出部分实体, 但整体识别结果不理想; 由于 CRF 模型对人工标记语料的信息量有较大依赖, 而 BiLSTM 模型可以融合上下文语义信息进行训练, 在标注语料信息较少时能够取得较好的识别结果, 故 CRF 模型的识别效果不及 BiLSTM 模型; TD-BiLSTM-CRF 模型则综合了 BiLSTM 模型和 CRF 模型的优点, 进一步提高了识别率. 分析 4 种模型的单个标签识别率, TD-BiLSTM-CRF 模型中的 PS 标签和 PR 标签识别率远高于其他 3 种模型. 这是由于 BiLSTM-CRF 模型不仅可以结合上下文信息, 而且还可以提取到句子前后标签间的依赖关系, 且对数字增加了特征标志, 把数字与字母连成的字段关联为一个切分单位, 而不是没有意义的数字与字母, 增强了后续 BiLSTM 层对句子特征的提取, 提升了工艺参数的区分和识别效果, 使得最终的 F_1 值达到 92%, 验证了本文模型的有效性.

此外, 4 种模型对 B-PF 标签的识别率均不高, 这可能是由于“外圆面”“面”“外圆柱面”“外圆”等相似特征的不同表述或者不同特征的相似表述, 影响了 BiLSTM 层的特征学习, 使得模型无法精准识别起始词.

(3) 不同数据集训练结果对比分析

为了判定该模型对不同文本的书写形式和书写习惯的鲁棒性, 本文选取了不同工艺手册中的典型零件加工工艺规程进行实验, 分别为数据集 1, 数据集 2 和数据集 3, 经过模型训练结果如表 4 所示.

从表 4 中可以看出, 在选取不同的数据集作为

表3 不同模型训练结果

%

标签	HMM ^[4]			CRF ^[3]			BiLSTM			TD-BiLSTM-CRF		
	准确率	召回率	F_1	准确率	召回率	F_1	准确率	召回率	F_1	准确率	召回率	F_1
B-PF	86.67	88.64	87.64	80.49	75.00	77.65	73.91	77.27	75.56	80.00	81.82	80.90
B-PA	90.91	89.20	85.11	92.59	100.00	96.15	96.15	100.00	98.04	96.15	100.00	98.04
B-PR	28.57	40.00	33.33	100.00	40.00	57.14	100.00	20.00	33.33	100.00	80.00	88.89
B-PM	95.45	100.00	97.67	97.56	95.24	96.39	95.35	97.62	96.47	93.35	97.62	96.47
B-PS	66.67	83.87	74.29	79.41	87.10	83.08	96.77	96.77	96.77	96.77	96.15	98.04
I-PF	87.76	93.48	90.53	92.68	82.61	87.36	94.87	80.43	87.06	90.70	84.78	87.64
I-PA	37.50	100.00	54.55	75.00	100.00	85.71	100.00	100.00	100.00	100.00	100.00	100.00
I-PR	30.00	66.67	41.38	100.00	66.67	80.00	100.00	66.67	80.00	100.00	100.00	100.00
I-PS	70.69	78.85	74.55	79.31	88.46	83.64	100.00	92.31	96.00	100.00	96.15	98.04
O	96.17	83.39	89.33	92.14	95.20	93.65	91.93	96.68	94.24	95.26	96.31	95.78
总体	88.00	84.85	85.87	87.00	89.20	88.01	89.09	91.92	89.98	90.20	93.88	92.00

表4 不同数据集训练结果

%

数据集	准确率	召回率	F_1
1	90.20	93.88	92.00
2	87.41	93.65	90.42
3	98.21	85.94	91.67

测试文本集进行实体识别时, TD-BiLSTM-CRF 算法仍然表现出优秀的性能, 并没有因为数据集的不同而产生显著差别, 这表明本文模型对不同数据集的书写形式具有鲁棒性。

4 结 语

本文通过分析机加工工艺规程文本的特点, 建立一种 TD-BiLSTM-CRF 识别模型, 实现了工步内容语料集中 5 类工艺信息的抽取。首先, 通过结合词典的 jieba 分词技术对数据集进行自动标注, 加强数字与标志字母之间的关联, 增强了 BiLSTM 神经网络的特征提取效果。其次, 经过 BiLSTM-CRF 进行训练输出最优标签序列。最后, 通过对比不同参数下的训练结果得出本文模型的 F_1 值达到 92%, 并与传统模型进行比较, 验证了本文模型的有效性。当选取不同的数据集作为测试文本集进行实体识别时, 本文模型仍表现出优秀的性能。未来工作将针对工艺规程文件开展关系抽取研究, 并与实体识别结合进行机加工领域知识图谱的构建。

参考文献(References):

- [1] Zhang Faping, Li Li. Research on knowledge push method for business process based on multidimensional hierarchical con-

text model[J]. Journal of Computer-Aided Design & Computer Graphics, 2017, 29(4): 751-758(in Chinese)

(张发平, 李丽. 基于多维层次情境模型的业务过程知识推送方法研究[J]. 计算机辅助设计与图形学学报, 2017, 29(4): 751-758)

- [2] Tang Jiayu, Liu Zhiyuan, Sun Maosong. A survey of text visualization[J]. Journal of Computer-Aided Design & Computer Graphics, 2013, 25(3): 273-285(in Chinese)
(唐家渝, 刘知远, 孙茂松. 文本可视化研究综述[J]. 计算机辅助设计与图形学学报, 2013, 25(3): 273-285)
- [3] Zhang Y J, Xu Z T, Zhang T. Fusion of multiple features for Chinese named entity recognition based on CRF model[C] //Proceedings of the 4th Asia Information Retrieval Symposium. New York: ACM, 2008: 95-106
- [4] Sun S C, Yun J, Lin H F, *et al.* Granular transfer learning using type-2 fuzzy HMM for text sequence recognition[J]. Neurocomputing, 2016, 214: 126-133
- [5] Collobert R, Weston J, Bottou L, *et al.* Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12(1): 2493-2537
- [6] Khalifa M, Shaalan K. Character convolutions for Arabic named entity recognition with long short-term memory networks[J]. Computer Speech & Language, 2019, 58: 335-346
- [7] Yu S, Cheng Y, Xie L, *et al.* A novel recurrent hybrid network for feature fusion in action recognition[J]. Journal of Visual Communication and Image Representation, 2017, 49(11): 192-203
- [8] Liu Yupeng, Li Dongdong. Chinese named entity recognition method based on Bi-directional LSTM-CNN-CRF[J]. Journal of Harbin University of Science and Technology, 2020, 25(1): 115-120(in Chinese)
(刘宇鹏, 栗冬冬. 基于 BiLSTM-CNN-CRF 的中文命名实体识别方法[J]. 哈尔滨理工大学学报, 2020, 25(1): 115-120)
- [9] Liu J X, Wu F Z, Wu C H, *et al.* Neural Chinese word segmentation with dictionary knowledge[C] //Proceedings of CCF International Conference on Natural Language Processing and Chinese Computing. Heidelberg: Springer, 2018: 80-91
- [10] Chen Yiren, He Kunjin, Lu Fengwei, *et al.* Discrimination of

- fracture types based on neural network and voxelized bone template[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(8): 1295-1307(in Chinese)
(陈义仁, 何坤金, 陆丰威, 等. 基于神经网络和体素模板的骨骼受损类型判别[J]. 计算机辅助设计与图形学学报, 2021, 33(8): 1295-1307)
- [11] Ji B, Liu R, Li S S, *et al.* A hybrid approach for named entity recognition in Chinese electronic medical record[J]. BMC Medical Informatics and Decision Making, 2019, 19(2): 149-158
- [12] Chhim P, Chinnam R B, Sadawi N. Product design and manufacturing process based ontology for manufacturing knowledge reuse[J]. Journal of Intelligent Manufacturing, 2019, 30(2): 905-916