高质量文本数据驱动的命名实体识别加速镍基单晶高温合金材料知识发现

刘 悦[1] 姚文轩[1] 刘大晖[1] 丁 琳[1] 杨正伟[1]
刘 微[2] 于 涛[3] 施思齐[2,4]

1 上海大学 计算机工程与科学学院 上海 200444
2 上海大学 材料基因组工程研究院 上海 200444
3 钢铁研究总院 功能材料研究所 北京 100081
4 上海大学 材料科学与工程学院 上海 200444

摘 要 镍基单晶高温合金构效关系知识常常以非结构化文本的形式存储在海量公开发表的科学文献中。利用命名实体识别(NER)方法从非结构化文本中挖掘关键信息已成为助力新材料研发的重要方式。然而，已有NER方法依赖于大量语料数据支撑且不适用于处理跨领域任务，导致其难以适配镍基单晶高温合金领域。本工作提出基于语义特征融合的深度学习命名实体识别方法(SF-NER)，以准确挖掘摘要文本中蕴含的镍基单晶高温合金知识。在领域知识指导下创建材料领域词典以实现远程监督，并建立了高质量镍基单晶高温合金标注语料库(含8类实体类型的19405个实体数据)；为准确捕捉特定材料术语，提出了融合编码的词表征策略以捕获关键材料语义特征；构建双向长短期记忆网络-条件随机场(Bi-LSTM-CRF)模型捕捉句子序列中的关键语义信息以实现实体标签的精准预测。实验结果表明，SF-NER能够精准识别镍基单晶高温合金实体类别(评价指标*F*1值为0.84)，有效筛选影响高温合金服役性能的关键因素，并推荐出可用于服役性能构效关系挖掘的高重要度描述符。

关键词 数据质量，深度学习，命名实体识别，镍基单晶高温合金，领域知识
中图分类号 TG131 **文章编号** 0412-1961(2024)10-1429-10

Named Entity Recognition Driven by High-Quality Text Data Accelerates the Knowledge Discovery of Nickel-Based Single Crystal Superalloys

LIU Yue[1], YAO Wenxuan[1], LIU Dahui[1], DING Lin[1], YANG Zhengwei[1], LIU Wei[2], YU Tao[3], SHI Siqi[2,4]

1 School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
2 Materials Genome Institute, Shanghai University, Shanghai 200444, China
3 Division of Functional Materials, Central Iron and Steel Research Institute, Beijing 100081, China
4 School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China

Correspondent: SHI Siqi, professor, Tel: 15800543880, E-mail: sqshi@shu.edu.cn

Supported by National Natural Science Foundation of China (Nos.52073169 and 92270124) and National Key Research and Development Program of China (No.2021YFB3802101)

Manuscript received 2024-06-11, in revised form 2024-07-09

资助项目 国家自然科学基金项目Nos.52073169和92270124，及国家重点研发计划项目No.2021YFB3802101
收稿日期 2024-06-11 **定稿日期** 2024-07-09

作者简介 刘 悦，女，1975年生，博士
通讯作者 施思齐，sqshi@shu.edu.cn，主要从事电化学储能材料计算与设计研究

DOI 10.11900/0412.1961.2024.00197

ABSTRACT The knowledge regarding the structure–activity relationships of nickel-based single crystal superalloys is mainly stored in the form of unstructured text in the vast published scientific literature, and its effective utilization can accelerate the design of high-performance materials. Named entity recognition (NER) methods can be used to extract vital information from unstructured text, thus contributing to automatically achieving tedious text mining tasks. However, existing NER methods typically rely on a large amount of corpus data, especially of the deep-learning-based type, and can hardly tackle cross-domain tasks. To the best of our knowledge, no prior research has been conducted for the knowledge discovery of nickel-based single crystal superalloys based on deep-learning-based NER; thus, it is difficult to adapt existing methods to this field. Here, a semantic-features-fused NER (SF-NER) method based on deep learning was proposed, aiming to accurately extract knowledge from abstract text concerning nickel-based single crystal superalloys. Specifically, as data quality determines the performance of NER models, a high-quality annotated corpus dataset for the above-mentioned superalloys (containing 19405 entity data of eight entity types) was constructed. This was created via remote supervision using domain-specific materials dictionary under the domain knowledge's guidance. To accurately capture the terms related to specific materials from the high-quality corpus dataset, a encoding fusion strategy for word representation was proposed for encoding the essential semantic features of materials from various perspectives. Then, based on these semantic features, a deep learning model, *i.e.*, bidirectional long short-term memory-cenditional random field (Bi-LSTM-CRF), was built to capture key semantic information in sentence sequences, thus accurately predicting entity types. The results of the experiment demonstrated that the proposed SF-NER method could accurately distinguish the entity categories of nickel-based single crystal superalloys (*i.e.*, $F1 = 0.84$) and effectively identify the key factors influencing their service performance. Lastly, descriptors with high importance were recommended, as they can be employed for machine learning modeling to explore the structure–activity relationships of high-performance materials.

KEYWORDS data quality, deep learning, named entity recognition, nickel-based single crystal superalloy, domain knowledge

材料科学文献是服役性能构效关系知识的主要载体。随着文献数量呈指数式增长，研究人员越来越难以通过人工检索获取助力高性能材料研发的高价值信息，且这种获取方式伴随着主观性和知识局限性等消极影响。因此，如何快速从文本中自动获取高价值信息是材料文本挖掘研究亟需解决的关键问题[1]。

命名实体识别(named entity recognition，NER)技术是信息提取的一个子任务，其目标是在文本中定位并分类预先定义的命名实体类别。因其在自动挖掘文本数据中关键信息的能力而被广泛应用于电池材料[2]、固态化学材料[3]、金属氧化物[4]以及无机材料[5]等领域。材料NER识别工作主要采用传统的NER方法(即基于规则和词典[6,7]、基于统计学的方法[8,9])和基于深度学习的NER方法(如基于卷积神经网络[10]、循环神经网络[11]、双向长短期记忆网络[12]和Transformer[13]等方法)。例如，Kuniyoshi等[14]基于序列标注和启发式规则自动识别和提取隐藏在科学文献中的材料合成过程，识别精度达到0.887。这些工作体现了传统NER方法所具有的高识别准确性，但其需要依赖专家经验，且对于训练数据中未出现的实体和特征组合识别效果不佳，因此不利于模型扩展和泛化。而基于深度学习的NER方法则能够有效缓解这些问题。例如，Kim等[4]利用Word2Vec词嵌入模型表示从1902~2018年间的50万份摘要中发掘出隐含领域知识并成功预测出尚未被发现的高性能新型热电材料。进一步，该团队还通过双向长短期记忆网络(Bi-directional long short-term memory，Bi-LSTM)和条件随机场(conditional random field，CRF)模型从327万篇材料科学文献摘要中提取出8000多万个命名实体。本课题组[15]提出了基于命名实体识别与文本数据增强的自动描述符识别方法，旨在从文本数据中实现嵌入领域知识的数据增强以及从粗粒度至细粒度对任务相关的描述符进行筛选，并在NASICON型固态电解质中得到很好的应用。

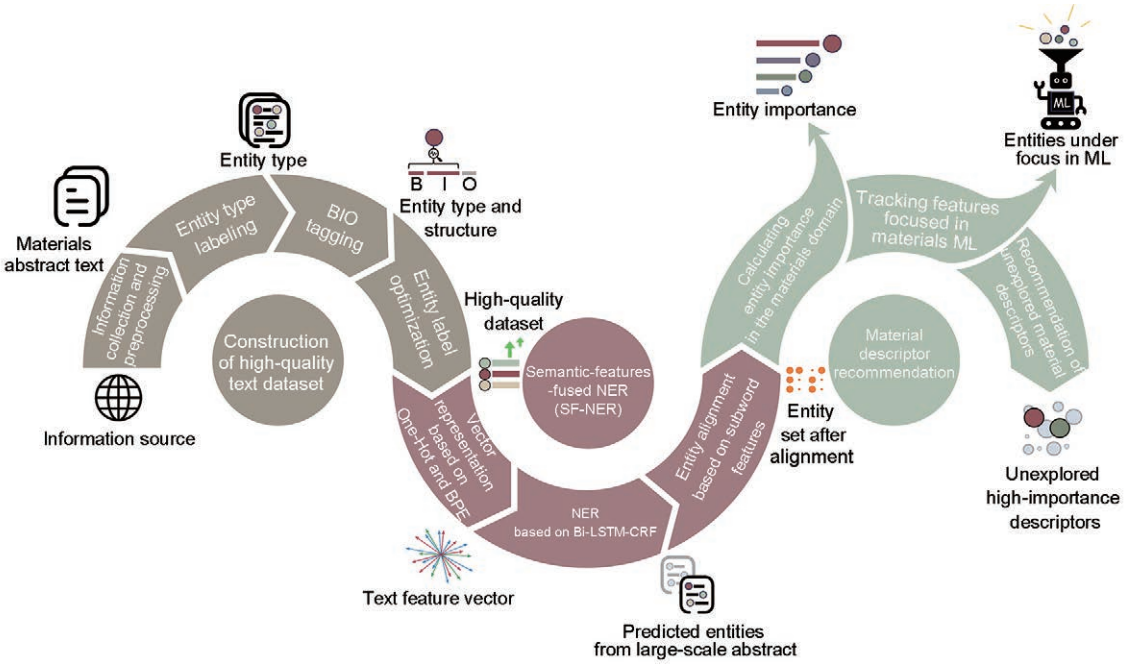
目前，NER在高温合金文献挖掘研究中仍处于起步阶段。例如，Sasidhar等[16]利用过程感知深度神经网络(DNN)模型整合耐腐蚀合金加工过程和电化学测试方法的文本数据，以及合金组成和环境测试参数的数值数据，成功对新设计的合金和测试条件进行坑蚀电位预测；Wang等[6]开发的材料数据提取工具基于预定义规则和启发式文本多关系提取算法实现，从14425篇文献中提取出2531个γ'相溶解温

度、密度、固相和液相线温度等钴基高温合金成分和性能数据，且与真实数据间的相对误差仅为0.81%；本课题组[17]提出了基于文本挖掘框架的知识发现方法，从来源时效性、发表权威性和作者学术地位等多维度角度对科学摘要的可信度进行量化，设计了8种描述镍基单晶高温合金的实体类型和领域词典，实现了从科学文摘中进行精准命名实体识别，最终通过分析获得了镍基单晶高温合金中重要化学成分的含量。然而，现有的基于深度学习的方法通常需要大量标注语料数据进行训练，且在处理跨领域任务时会因难以识别专业术语而迁移效果不佳。同时，其识别效果依赖于监督语料数据的质量[18−20]，使得数据集构建方式至关重要。因此，如何在领域知识的指导下[21,22]构建适用于镍基单晶高温合金的高质量有标注语料数据，并据此构建针对性的深度学习NER方法是亟需解决的难点问题。

因此，本工作提出基于语义特征融合的深度学习命名实体识别(SF-NER)方法，以准确提取镍基单晶高温合金材料文本中的实体。首先，在领域知识指导下，预定义了8种材料实体类型，并结合人工标注与基于领域词典的远程标注方法，构建了高质量的镍基单晶高温合金标注语料。其次，为准确捕捉特定材料术语，提出融合独热编码(One-Hot encoding)和字节对编码(BPE)表示的词表征方式，并耦合Bi-LSTM和CRF模型对句子序列进行建模和标签预测。然后，基于子词向量对材料实体进行同义词对齐，并结合文献可信度和词频-逆文档频率(TF-IDF)对统一后的实体进行重要度分析。最后，在领域专家指导下，通过对比高重要度的实体和材料机器学习研究中已使用的描述符，验证该方法对于挖掘材料描述符的有效性，从而为机器学习建模推荐合适的特征。该方法可以加速从文本中发现镍基单晶高温合金材料知识并有望推广至其他材料领域。

1 方法

基于SF-NER方法进行镍基单晶高温合金文献挖掘及应用流程如图1所示。首先，以镍基单晶高温合金材料文献摘要文本为基础，经过文本预处理、实体类型标注、BIO标注(其中“B”表示实体的开头(Begin)，“I”表示实体的中间部分(Inside)，“O”表示非实体(Outside))和实体标签优化等步骤，构建适用于NER任务的高质量文本数据集；基于此，训练融合语义特征的Bi-LSTM-CRF模型，耦合One-Hot[23]与BPE编码[24]对材料领域文本进行特征融合的向量化表示，并通过Bi-LSTM和CRF模型对语句序列进行标签预测，利用子词级特征对同义实体进行对齐，从而得到镍基单晶高温合金材料实体集；最后，以镍基单晶高温合金材料描述符挖掘与推荐为例，探索



Color online

图1 基于语义特征融合的深度学习命名实体识别(SF-NER)方法的镍基单晶高温合金文献挖掘及应用流程图

Fig.1 Diagram for literature mining and field application of nickel-based single crystal superalloys using the SF-NER (BPE—byte-pair encoding, NER—named entity recognition, ML—machine learning, CRF—conditional random field, B—the beginning of an entity (Begin), I—the inside of an entity (Inside), O—outside of an entity (Outside))

本文方法在材料领域研究中的应用，即结合文献可信度[17]和TF-IDF方法[25]对统一后的实体进行重要度分析，并在材料领域专家指导下进行机器学习建模用描述符推荐，从而达到对镍基单晶高温合金材料的构效关系知识的挖掘与应用。

1.1 高质量文本数据集构建

高质量的材料文本数据集是实现高精度材料文本挖掘的基础。因此，本工作首先开展了适用于镍基单晶高温合金材料NER任务的高质量科学文本数据集构建研究。具体包含3个阶段：材料科学文献信息采集与预处理、基于预定义实体类型的材料文本标注和融合领域知识的材料实体标签优化。

在材料科学文献信息采集与预处理阶段，以“nickel base single crystal superalloy”及其同义短语为检索主题，从Web of Science数据库中收集了631篇摘要，并采用NLTK (Natural Language Toolkit)[26]工具对摘要文本进行分词和词性标注以完成预处理。

在基于预定义实体类型的材料文本标注阶段，基于文献[17]中定义的涵盖镍基单晶高温合金领域实体的8种材料实体类型，明确需标注材料实体的含义和描述。其具体定义和详细描述如定义1和补充材料表S1所示。基于此，根据文献[18]中的分析，选择适配于实体标注任务且可扩展性高的EasyData标注工具开展多人联合实体标注，从而快速得到具有实体类型的语料。进一步，为了更加清晰地界定材料实体边界位置，使用ChemDataExtractor工具包对语料数据进行BIO标注[27]，将文本标签格式化为“B-实体类型”、“I-实体类型”和“O”的形式。

定义1：材料实体类型= <APL, CMT, CMP, CNT, FAT, PRO, PRP, STE>。其中APL、CMT、CMP、CNT、FAT、PRO、PRP、STE分别代表应用(Application)、表征(Characterization)、成分(Composition)、条件(Condition)、特征(Feature)、加工(Processing)、属性(Property)和结构(Structure)，是用来描述和表达材料信息的特殊术语。

在融合领域知识的材料实体标签优化阶段，针对基于人工标注的方法存在标注不一致与标签不均衡的问题，本工作采用基于领域词典的远程监督标注方法[17]自动化地对摘要文本再次进行标注。该方法由词典中的领域知识驱动，能够捕捉材料特征明显的实体。进一步，在领域专家的指导下整合2种标注结果不一致的部分，从而确保标注数据的准确性和覆盖率，提升标注数据的质量。最终，形成了基于人工标注的文本数据集A_ManualLabeling和优化后的高质量文本数据集A_DomainDictionary。详细信息见补充材料表S2。

1.2 融合语义特征的Bi-LSTM-CRF命名实体识别

1.2.1 基于One-Hot与BPE的材料文本向量表示 One-Hot编码通过将每个单词映射为一个高维稀疏向量实现词表示，其中向量的维度与词汇表的规模相等，且仅有一个元素为1，其余为0。这种表示方法直观简洁，但在处理大规模语料库时容易导致维度灾难。BPE编码采用子词级别的策略，通过迭代合并语料库中频繁出现的字符序列生成子词，并将材料文本表示为这些子词的序列，有效缓解了未登录词和稀有词问题，同时降低了向量的维度。

为了兼顾One-Hot编码的离散特性和BPE编码的词级别的表示优势，本工作提出一种融合策略。具体而言，每个单词首先通过One-Hot编码转换为高维稀疏向量。然后，使用BPE编码对单词进行子词分解，并将每个子词映射为固定维度的向量。最后，通过拼接单词的One-Hot向量与其子词向量，形成融合词向量。这种融合向量不仅捕捉了单词的语法和语义信息，还增强了模型对未登录词和稀有词的处理能力，从而提升命名实体识别的准确性。材料文本向量融合过程如式(1)~(3)所示：

$$x_i^{\mathrm{OneHot}} = \mathrm{OneHot}\left(\omega_i\right) \tag{1}$$

$$x_i^{\mathrm{BPE}} = \mathrm{BPE}\left(\omega_i\right) \tag{2}$$

$$x_i = \left[x_i^{\mathrm{OneHot}}; x_i^{\mathrm{BPE}}\right] \tag{3}$$

式中，ω_i为词汇表中第i个词。x_i^{OneHot}为词汇ω_i的One-Hot编码向量。x_i^{BPE}为词汇ω_i的BPE编码向量。x_i为将这2种向量拼接的结果。

1.2.2 基于Bi-LSTM-CRF的命名实体识别 为缓解机器学习在执行跨领域任务时迁移效果不佳的问题，本工作采用基于深度学习的命名实体识别方法Bi-LSTM-CRF进行实体标签预测。Bi-LSTM-CRF模型是一种常用于序列标注任务的深度学习模型。该模型结合Bi-LSTM和CRF模型，旨在通过学习上下文信息和标记序列之间的依赖关系来提高序列标注的性能。

首先，使用Bi-LSTM网络学习材料文本序列在前后2个方向上的信息。将融合One-Hot与BPE 2种编码形式的词向量输入到Bi-LSTM中，通过双向遍历文本序列同时捕捉到单词前后的上下文信息。然后，将2个方向上的序列处理结果进行拼接。其中，Bi-LSTM在t时刻隐藏层状态h_t的更新过程如式

(4)~(6)所示:

$$\overrightarrow{\boldsymbol{h}_t} = \mathrm{LSTM}_{\mathrm{forward}}(x_t, \overrightarrow{\boldsymbol{h}_{t-1}}) \tag{4}$$

$$\overleftarrow{\boldsymbol{h}_t} = \mathrm{LSTM}_{\mathrm{backward}}(x_t, \overleftarrow{\boldsymbol{h}_{t+1}}) \tag{5}$$

$$h_t = \left[\overrightarrow{\boldsymbol{h}_t}; \overleftarrow{\boldsymbol{h}_t}\right] \tag{6}$$

式中,x_t为在t时刻的输入向量;$\overrightarrow{\boldsymbol{h}_t}$为前向LSTM在$t$时刻的隐藏状态;$\overleftarrow{\boldsymbol{h}_t}$为后向LSTM在$t$时刻的隐藏状态;$h_t$为在$t$时刻的隐藏状态,由前向和后向隐藏状态拼接而成,包含了在t时刻上下文的双向信息。

其次,利用CRF模型对Bi-LSTM层的输出进行解码,计算最优的标签序列。CRF考虑了序列中相邻标签之间的依赖关系,可以有效地捕获标签之间的转移特征,并通过联合概率分布对标签序列进行全局归一化,从而提高序列标注的准确性。假设给定输入序列的Bi-LSTM输出序列为$h=(h_1, h_2, \cdots, h_T)$,其中$h_i$是第$i$ ($i=1, 2, \cdots, T$)个单词的输出向量,$y=(y_1, y_2, \cdots, y_T)$表示对应于$h$的标签序列,CRF模型的标注得分score($h$, y)由式(7)计算:

$$\mathrm{score}\left(h, y\right) = \sum_{t=1}^{T}\left(\boldsymbol{W}^T\left[h_t; y_{t-1}\right] + b_{\mathrm{tag}}\right) + \boldsymbol{E}\left[y_T\right] \tag{7}$$

式中,$\boldsymbol{W}$是权重矩阵;b_{tag}是标签偏置向量;$\boldsymbol{E}$是发射矩阵,由Bi-LSTM层计算得到;y_{t-1}是$t-1$时刻的标签;$\left[h_t; y_{t-1}\right]$是$t$时刻Bi-LSTM的输出和上一时刻CRF的标签的拼接向量。

最后,句子S基于Bi-LSTM-CRF模型生成标签序列y的概率$P(y|S)$由式(8)计算:

$$P\left(y|S\right) = \frac{\exp\left(\mathrm{score}\left(h, y\right)\right)}{\sum_{\tilde{y}\in y}\exp\left(\mathrm{score}\left(h, \tilde{y}\right)\right)} \tag{8}$$

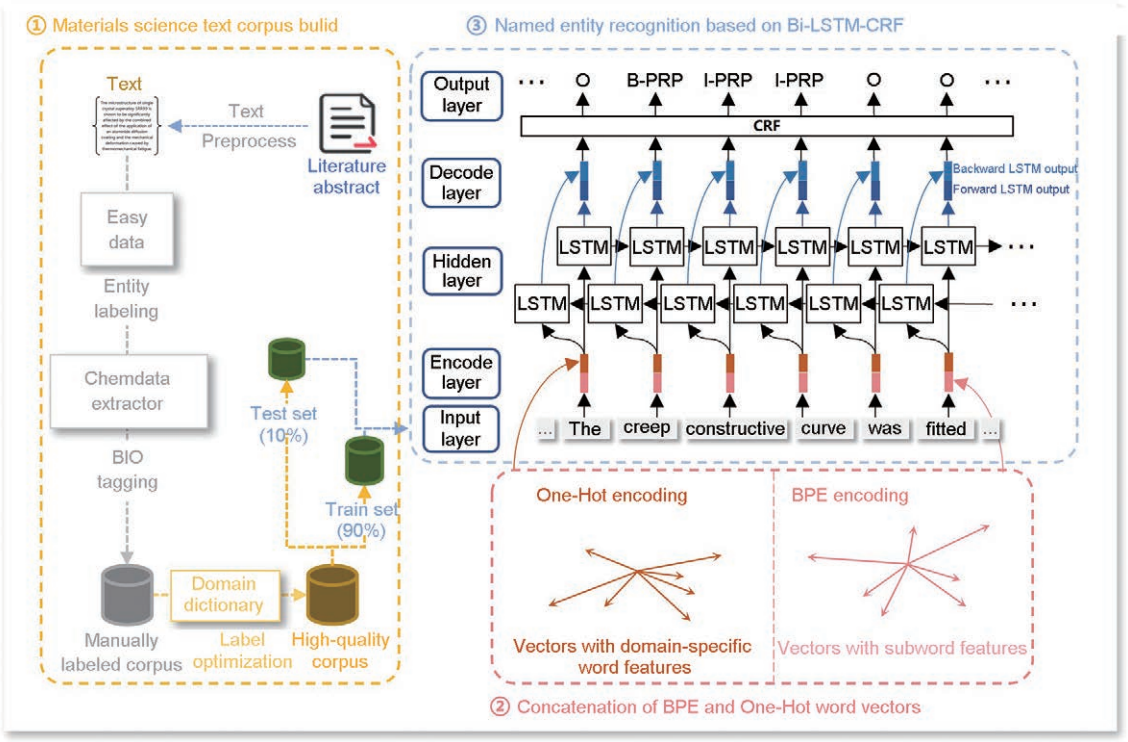
式中,$\tilde{y}$为真实标签。

在训练过程中,Bi-LSTM-CRF模型通过最小化负对数似然损失来优化参数,对数似然函数如式(9)所示。在模型训练完成后,使用Viterbi算法[28]找到最佳的标签序列。Viterbi算法通过递推计算每个时刻每个状态的最大概率路径,从而获得如式(10)所示的最优标签序列y^*。融合语义特征的Bi-LSTM-CRF命名实体识别方法流程与框架如补充材料算法S1与图2所示。

$$\ln\left(P\left(y|S\right)\right) = \mathrm{score}\left(h, y\right) - \ln\left(\sum_{\tilde{y}\in y}\exp(\mathrm{score}(h, \tilde{y}))\right) \tag{9}$$

$$y^* = \arg\max_{\tilde{y}\in y}\ \mathrm{score}(h, \tilde{y}) \tag{10}$$

1.2.3 基于子词特征的材料实体统一　在镍基单晶高温合金领域,统一术语的多样表达形式和书写习惯是一项关键任务。为有效合并同义实体,本



Color online

图2 基于双向长短期记忆网络-条件随机场(Bi-LSTM-CRF)模型的材料命名实体识别框架

Fig.2 Material named entity recognition framework for Bi-LSTM-CRF (Bi-LSTM—bi-directional long short term memory, PRP—property)

工作采用基于BPE的方法，该方法通过深入分析材料实体词的子词特征，有助于识别实体间的共性，进而实现同义实体的有效合并。具体而言，首先通过BPE编码生成每个实体的嵌入向量，以捕捉实体的语义特征。然后，通过计算实体嵌入向量间的余弦相似度，评估实体间的语义相似性。进一步，基于设定的相似度阈值，判断是否将2个实体归类为同一类别。最后，采用并查集数据结构处理实体集群的合并与查询操作，通过并查集中的“查找”和“合并”功能，将相似度超过阈值的实体合并至同一集群，完成实体统一。该方法的流程如补充材料算法S2所示。

2 实验

2.1 实验设置

本工作模型的参数设置详见补充材料表S3。为了对比不同模型的性能，采用准确率(Precision)、召回率(Recall)及$F1$值($F1$-score)作为评价指标，计算过程如式(11)~(13)所示，指标值越大则模型的性能越好。其中$F1$值是准确率和召回率的调和平均数，是性能评估的关键指标。TP为预测正确的实体数量，FP为预测结果中错误的实体数量，FN为实际结果中未被预测正确的实体数量。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$F1\text{-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

2.2 实验结果与分析

2.2.1 SF-NER模型性能验证 为了验证SF-NER模型中的基于Bi-LSTM-CRF框架的命名实体识别方法及融合词向量表示方法的有效性，本工作对比了SF-NER与其消融模型以及BERT-Bi-GRU-CRF模型[29] (BERT—bi-directional encoder representations from transformers，GRU—gated recurrent unit)在数据集A_DomainDictionary上的表现，实验结果如表1所示。

在表1中，SF-NER模型在准确率、召回率和$F1$值上均优于BERT-Bi-GRU-CRF模型，达到了84%，证明了Bi-LSTM-CRF在处理镍基单晶高温合金材料文本数据集上的有效性，这归因于Bi-LSTM能够有效捕捉输入序列的上下文语义特征，而以Encoder为基础的BERT模型在执行特定领域任务时高度依赖于其预训练过程，且相较于传统深度学习模型而言，在适应下游任务时对数据量的要求更高。此外，

表1 不同模型在数据集A_DomainDictionary上的准确率、召回率和$F1$值

Table 1 Precision, recall, and $F1$-score of different models on dataset A_DomainDictionary

Model	Precision	Recall	$F1$-score
BERT-Bi-GRU-CRF	0.57	0.62	0.60
Bi-LSTM(Glove)-CRF	0.80	0.81	0.80
Bi-LSTM(OneHot)-CRF	0.81	0.81	0.81
Bi-LSTM(OneHot-Glove)-CRF	0.82	0.83	0.82
SF-NER	0.84	0.84	0.84

Note: BERT-Bi-GRU-CRF—bi-directional encoder representations from transformers (BERT) -bidirectional gated recurrent unit (GRU)-CRF

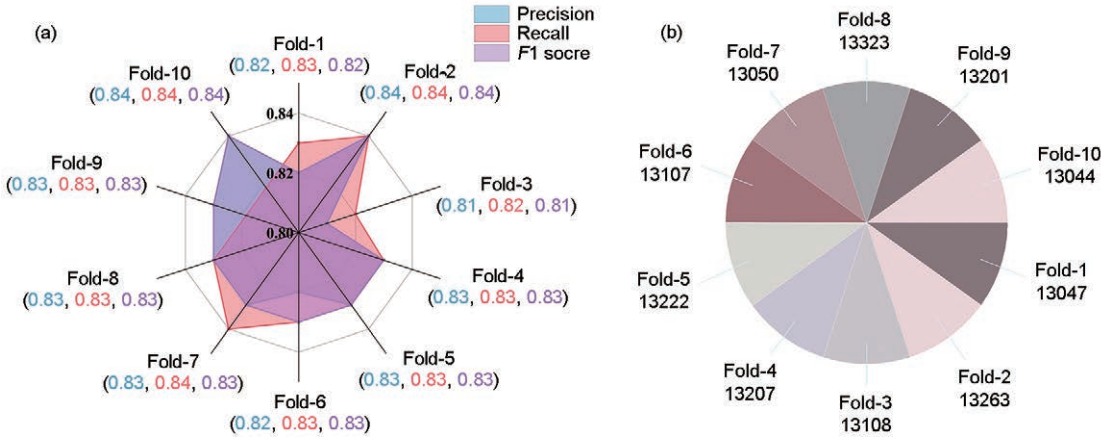
通过特征编码消融实验，本工作探讨了编码策略对模型性能的影响。由于BPE编码词向量无法单独应用于材料NER模型，采用Glove[30]编码模型作为平替。通过比较Bi-LSTM(OneHot-Glove)-CRF模型与仅使用Glove或One-Hot的Bi-LSTM-CRF模型表现，发现前者在各项指标上均表现更佳，从而验证了词向量融合策略的有效性。进一步地，SF-NER相较于Bi-LSTM(OneHot-Glove)-CRF模型在准确率、召回率和$F1$值上均提升了2%左右。这是因为融合BPE和One-Hot的表征方法能够表示子词级别的词向量，使得Bi-LSTM网络充分学到特定领域词特征，从而提升模型对文本标签的学习和预测能力。

为了进一步说明SF-NER模型的稳定性，本工作在数据集A_DomainDictionary上进行了十折交叉验证，图3展示了模型在十折交叉验证时的性能表现和识别的实体个数。由图3a可知，模型的准确率、召回率和$F1$值均能保持在0.80以上。此外，由图3b可知，每一折交叉验证均基于超过13000个词实体进行。即使在处理大规模数据时，模型依然表现出稳定的性能，这体现了该模型在真实应用场景中的实际预测能力和适应性，从而保障了其在下游文本挖掘任务中的应用效果。

2.2.2 数据集质量验证 为了评估基于领域词典的远程标注方法对于文本数据集质量方面的提升效果，对比了不同模型在数据集A_ManualLabeling以及A_DomainDictionary上的表现，实验结果如表2所示。结果显示，所有模型在质量优化后的数据集A_DomainDictionary上的表现显著优于在人工标注数据集A_ManualLabeling上的表现。特别是在基准模型BERT-Bi-GRU-CRF上，$F1$差值高达0.16。

这表明，A_DomainDictionary 数据集具备更高的质量；结合领域知识驱动的远程标注方法能够有效构建高质量材料文本数据集，从而显著提高模型的预测性能。

2.2.3 材料实体识别结果分析　进一步分析 SF-NER 模型对不同类型实体的识别效果，其表现如图 4 所示。其中，SF-NER 模型对于“Condition”类实体的识别效果最好，*F*1 值达到 80%；而对于“Feature”类实体的识别效果最差，*F*1 值仅有 32%，这一表现主要源于“Feature”类型实体在文本中的边界模糊性。例如，“single-crystal”作为独立实体或与“alloy”结合形成“sing-crystal alloy”均可被视作为正确识别，但此类边界模糊性增加了识别的不确定性，从而影响了模型对于该类别实体的识别效果。为了



Color online

图 3 SF-NER 模型在数据集 A_DomainDictionary 上的十折交叉验证结果

Fig.3 Ten-fold cross-validation results of the SF-NER on dataset A_DomainDictionary

(a) performance of SF-NER during ten-fold cross-validation

(b) number of word entities recognized by SF-NER during ten-fold cross-validation

表 2 不同模型在 2 个数据集上的表现(*F*1 值)

Table 2 Performance (*F*1-score) of different models on two datasets

Model	A_ManualLabeling	A_DomainDictionary
BERT-Bi-GRU-CRF	0.44	0.60
Bi-LSTM(Glove)-CRF	0.75	0.80
Bi-LSTM(OneHot-BPE)-CRF	0.78	0.84

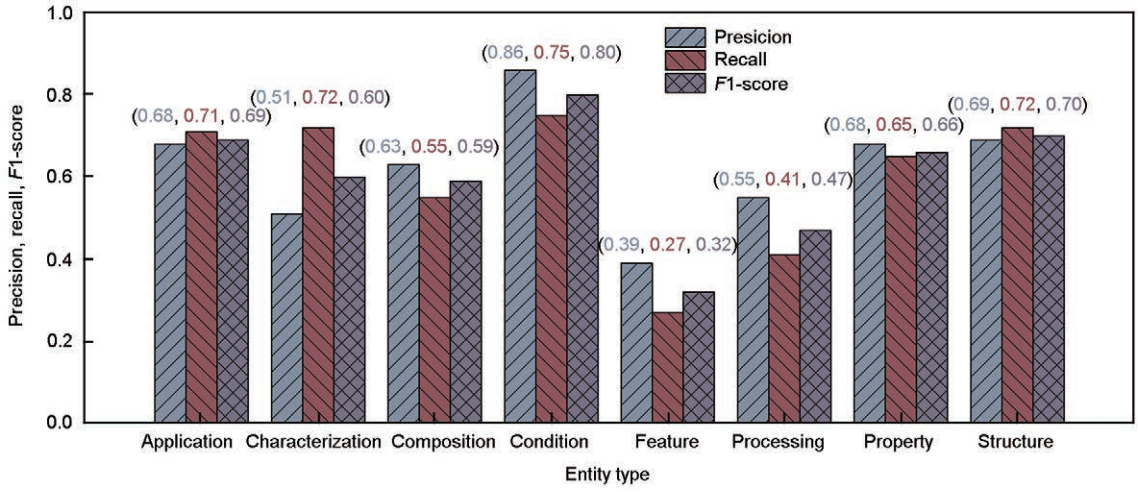


图 4 SF-NER 模型对不同类型实体的识别准确率、召回率和 *F*1 值

Fig.4 Precision, recall, and *F*1-score of the SF-NER model for different-type entities

避免这种影响，可考虑更严格地定义“Feature”类型实体的边界。

3 应用探索

本工作将可用描述符的挖掘与推荐作为应用实例用于验证SF-NER方法进行知识挖掘的有效性。具体地，首先采用SF-NER模型从大规模摘要文本中抽取实体，并结合文献可信度评估方法和TF-IDF方法计算实体的重要度；其次，将实体的重要度作为推荐系数，并进行推荐系数排名与描述符推荐。最终，在材料领域专家的指导下，对具有高重要度的实体进行影响力分析，探讨此类实体对镍基单晶高温合金研究的促进作用。

3.1 实体重要度计算

为了提供描述符推荐所需的数据基础，采用2.1节中所述的数据采集方法，从Web of Science数据库中收集了1996年~2024年1月期间的5020篇有关镍基单晶高温合金材料的科学文献摘要，并对其进行预处理，以形成大规模的待挖掘摘要文本库。然后，使用2.1节中基于631篇摘要构建的高质量文本数据集训练SF-NER模型，用以从待挖掘的摘要文本库中识别材料实体，并对所识别的实体进行实体统一处理。最后，结合文献可信度评估方法与TF-IDF技术，对SF-NER模型识别出的实体进行重要度(I_E)计算。其计算公式如式(14)所示：

$$I_E = \frac{1}{|D|}\sum_{d_i \in D} C(E)\cdot \mathrm{TFIDF}(E, d_i, D) \qquad (14)$$

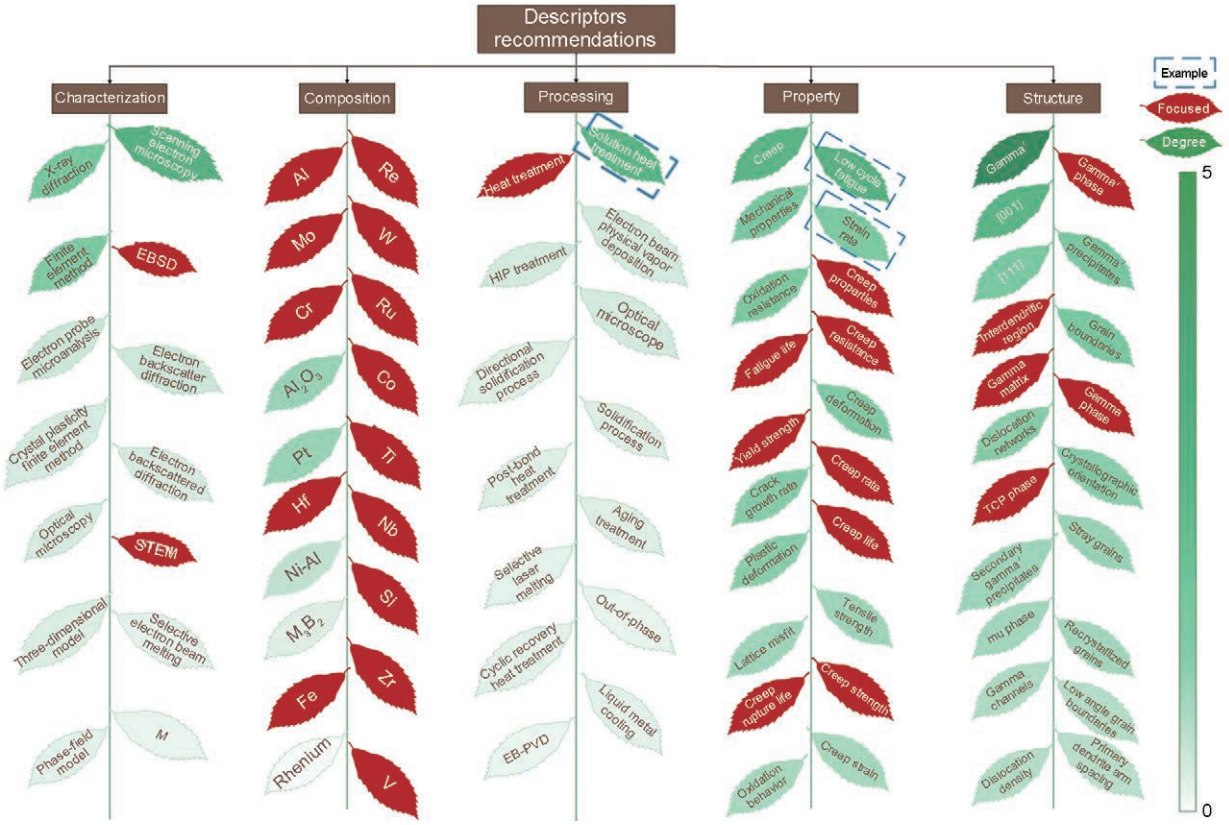
式中，D为镍基单晶高温合金材料领域语料库中的摘要集合，$|D|$为语料库中摘要的总数，$C(E)$为实体所属文献的可信度，d_i为摘要集合中的第i篇摘要。$\mathrm{TFIDF}(E, d_i, D)$为实体在摘要$d_i$中的TF-IDF值。由于TF-IDF方法用于衡量实体在语料库里某一文档中的重要程度，而文献可信度评估则反映了文献本身的可信程度，因此，实体重要度作为这2者的综合指标，能够真实可信地评估实体在镍基单晶高温合金材料领域中的重要性。从摘要文本中提取的各类型镍基单晶高温合金材料实体及其重要度如补充材料表S4所示。

3.2 基于重要度的描述符推荐

为了探究影响镍基单晶高温合金材料性能的关键因素，本工作进行了描述符推荐研究。首先，据补充材料表S4可知，在8个实体类型中，“Characterization”、“Composition”、“Processing”、“Property”和“Structure”类型的实体多为粗粒度可推广的实体。因此，为了获得抽象且可用性强的描述符，本工作将推荐描述符的候选来源范围限定在上述5类实体中。

接着，以重要度作为推荐系数，从各类型实体中选取推荐系数不低于0.19的共385个实体作为推荐描述符。同时，在材料领域专家的指导下，从22篇基于机器学习的镍基单晶高温合金材料构效关系研究文献中统计出95个已用于机器学习建模的描述符，具体如补充材料表S5所示。为研究各类型描述符的推荐程度及其在机器学习建模中的实际应用情况，对各类型的描述符以推荐系数进行降序排序，并统计了本工作推荐的385个描述符与95个已用于机器学习建模的描述符之间的重叠情况，如图5所示(部分展示)。通过对比发现，这95个描述符中有高达82个(图中红色叶子节点)为本工作推荐的描述符，且在各类型描述符中均排名靠前。这些描述符是能够有效预测和分析影响镍基单晶高温合金服役性能的关键因素，其高被推荐率及推荐高排名验证了本方法可以有效地为机器学习建模推荐描述符。

此外，图中排名较高(图中绿色叶子的颜色深浅及叶子高低对应推荐系数高低)的绿色叶子节点指代的描述符虽不在82个已用于机器学习建模的描述符之中，但其仍具备高重要度，其具体重要度数值见补充材料表S6，有为未来材料机器学习建模所用的潜力。例如，在图5虚线框住的示例中，重要度为1.9的“strain rate”(应变率)实体，其可能对合金材料的塑性和断裂行为的预测有帮助。因为高应变率通常会导致材料内部应力集中和变形局部化现象加剧，从而降低材料的塑性变形能力和抗断裂能力。实体重要度为2.62的“solution heat treatment”(固溶热处理)是一种通过控制合金在高温下固溶体的形成和溶解来改善材料性能的热处理工艺。固溶热处理过程中的参数可能是用于预测合金材料硬度的潜在因子，因为在固溶热处理过程中，合金中的固溶体溶解均匀，随后通过快速冷却或固溶体再结晶，可以获得细小、均匀的晶粒结构，从而显著提高了材料的硬度。类似地，实体重要度为2.94的“low cycle fatigue”(低周疲劳)指材料在较低的应变振幅下经历的疲劳过程。低周疲劳数据可能利于对合金材料残余寿命的精确预测，因为在低应变振幅下，重复加载会逐渐积累微观损伤和位错，导致材料强度和耐久性下降，同时增加疲劳裂纹形成的风险，进一步缩短材料的残余寿命。



Color online

图5 推荐的描述符及已被材料机器学习关注的实体(部分展示)

Fig.5 Recommended descriptor importance ranking and entities already focused on in materials machine learning (partial display) (HIP—hot isostatic pressing, TCP—topologically close-packed phases, EB-PVD—electron beam-physical vapor deposition, SRZ—secondary reaction zone, IDZ—interdiffusion zone)

综上所述，本研究可以从海量材料科学文本中自动获取潜在的高重要度描述符。材料专家可从中遴选并加工出可信、可靠的描述符作为机器学习建模的输入特征，有望进一步推动对镍基单晶高温合金服役性能与复杂微观内禀特性的深入挖掘。

4 结论

(1) 提出基于语义特征融合的深度学习命名实体识别方法 SF-NER。融合 One-Hot 和 BPE 的词表征方式以捕捉关键领域术语语义，并设计 Bi-LSTM-CRF 模型准确识别目标实体类型。该方法在高质量镍基单晶高温合金语料库上对 8 类材料实体识别的 *F*1 值达到 0.84，优于其他基于深度学习的 NER 方法。

(2) 构建了面向镍基单晶高温合金的高质量语料库。定义了应用、表征、成分、条件、形貌、工艺、性能、结构等 8 种材料实体类型，利用基于领域词典的高质量材料科学文本挖掘数据集构建方法从 631 篇镍基单晶高温合金材料摘要中标注了涵盖 8 类材料实体类型的 19405 个实体。

(3) 在领域专家指导下从 5020 篇镍基单晶高温合金材料摘要文本中挖掘并推荐出精准刻画“材料成分、工艺、组织、性能、服役行为”构效关系的高重要度描述符，有望进一步推动对镍基单晶高温合金服役性能构效关系的深入挖掘。研究人员可根据实际需求对实体类型进行更新与调整，以确保模型能够准确捕捉领域的实体类别信息，从而实现本工作方法有效扩展至其他材料领域的知识发现。

文中补充材料可通过以下网址查看：https://www.ams.org.cn/CN/10.11900/0412.1961.2024.00197

参 考 文 献

[1] Shi S Q, Tu Z W, Zou X X, et al. Applying data-driven machine learning to studying electrochemical energy storage materials [J]. Energy Storage Sci. Technol., 2022, 11: 739
(施思齐, 涂章伟, 邹欣欣等. 数据驱动的机器学习在电化学储能材料研究中的应用 [J]. 储能科学与技术, 2022, 11: 739)

[2] El-Bousiydy H, Lombardo T, Primo E N, et al. What can text mining tell us about lithium-ion battery researchers' habits? [J]. Batter. Supercaps, 2021, 4: 758

[3] Mahbub R, Huang K, Jensen Z, et al. Text mining for processing

conditions of solid-state battery electrolytes [J]. Electrochem. Commun., 2020, 121: 106860

[4] Kim E, Huang K, Saunders A, et al. Materials synthesis insights from scientific literature via text extraction and machine learning [J]. Chem. Mater., 2017, 29: 9436

[5] Huo H Y, Rong Z Q, Kononova O, et al. Semi-supervised machine-learning classification of materials synthesis procedures [J]. npj Comput. Mater., 2019, 5: 62

[6] Wang W R, Jiang X, Tian S H, et al. Automated pipeline for superalloy data by text mining [J]. npj Comput. Mater., 2022, 8: 9

[7] Hawizy L, Jessop D M, Adams N, et al. ChemicalTagger: A tool for semantic text-mining in chemistry [J]. J. Cheminf., 2011, 3: 17

[8] Leaman R, Wei C H, Lu Z Y. tmChem: A high performance approach for chemical named entity recognition and normalization [J]. J. Cheminf., 2015, 7: S3

[9] Kim E, Huang K, Jegelka S, et al. Virtual screening of inorganic materials synthesis parameters with deep learning [J]. npj Comput. Mater., 2017, 3: 53

[10] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition [J]. Neural Comput., 1989, 1: 541

[11] Williams R J, Zipser D. A learning algorithm for continually running fully recurrent neural networks [J]. Neural Comput., 1989, 1: 270

[12] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Comput., 1997, 9: 1735

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

[14] Kuniyoshi F, Makino K, Ozawa J, et al. Annotating and extracting synthesis process of all-solid-state batteries from scientific literature [A]. Proceedings of the 12th Language Resources and Evaluation [C]. Marseille: European Language Resources Association, 2020: 1941

[15] Liu Y, Ge X Y, Yang Z W, et al. An automatic descriptors recognizer customized for materials science literature [J]. J. Power Sources, 2022, 545: 231946

[16] Sasidhar K N, Siboni N H, Mianroodi J R, et al. Enhancing corrosion-resistant alloy design through natural language processing and deep learning [J]. Sci. Adv., 2023, 9: eadg7992

[17] Liu Y, Ding L, Yang Z W, et al. Domain knowledge discovery from abstracts of scientific literature on nickel-based single crystal superalloys [J]. Sci. China Technol. Sci., 2023, 66: 1815

[18] Liu Y, Liu D H, Ge X Y, et al. A high-quality dataset construction method for text mining in materials science [J]. Acta Phys. Sin., 2023, 72: 070701
(刘 悦, 刘大晖, 葛献远等. 高质量的材料科学文本挖掘数据集构建方法 [J]. 物理学报, 2023, 72: 070701)

[19] Liu Y, Ma S C, Yang Z W, et al. A data quality and quantity governance for machine learning in materials science [J]. J. Chin. Ceram. Soc., 2023, 51: 427
(刘 悦, 马舒畅, 杨正伟等. 面向材料领域机器学习的数据质量治理 [J]. 硅酸盐学报, 2023, 51: 427)

[20] Liu Y, Yang Z W, Zou X X, et al. Data quantity governance for machine learning in materials science [J]. Natl. Sci. Rev., 2023, 10: nwad125

[21] Liu Y, Zou X X, Yang Z W, et al. Machine learning embedded with materials domain knowledge [J]. J. Chin. Ceram. Soc., 2022, 50: 863
(刘 悦, 邹欣欣, 杨正伟等. 材料领域知识嵌入的机器学习 [J]. 硅酸盐学报, 2022, 50: 863)

[22] Shi S Q, Sun S Y, Ma S C, et al. Detection method on data accuracy incorporating materials domain knowledge [J]. J. Inorg. Mater., 2022, 37: 1311
(施思齐, 孙拾雨, 马舒畅等. 融合材料领域知识的数据准确性检测方法 [J]. 无机材料学报, 2022, 37: 1311)

[23] Goldberg Y. A primer on neural network models for natural language processing [J]. J. Artif. Intell. Res., 2016, 57: 345

[24] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. J. Artif. Intell. Res., 2011, 12: 2493

[25] Jones K S. A statistical interpretation of term specificity and its application in retrieval [J]. J. Doc., 1972, 28: 11

[26] Bird S. NLTK: The natural language toolkit [A]. Proceedings of COLING/ACL 2006 Interactive Presentation Sessions [C]. Sydney: Association for Computational Linguistics, 2006: 69

[27] Nadkarni P M, Ohno-Machado L, Chapman W W. Natural language processing: an introduction [J]. J. Am. Med. Inform. Assoc., 2011, 18: 544

[28] Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm [J]. IEEE Trans. Inform. Theory, 1967, 13: 260

[29] Lv J H, Du J P, Zhou N, et al. BERT-BIGRU-CRF: A novel entity relationship extraction model [A]. 2020 IEEE International Conference on Knowledge Graph [C]. Nanjing: IEEE, 2020: 157

[30] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation [A]. Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing [C]. Doha: Association for Computational Linguistics, 2014: 1532

(责任编辑:李海兰)