◆自然语言理解◆

基于BERT的危险化学品命名实体识别模型*

陈观林[1,2**],程　钊[1,2],邹　凌[2],杨武剑[1],李　甜[1]
(1.浙大城市学院,计算机与计算科学学院,浙江杭州　310015;2.常州大学计算机与人工智能学院,江苏常州　213164)

摘要:针对危险化学品实体识别及关系识别的问题,本文基于双向长短期记忆网络连接条件随机场(Bidirectional Long Short-Term Memory with Conditional Random Field,BiLSTM-CRF)模型,通过引入双向编码器表示(Bidirectional Encoder Representation from Transformers,BERT)模型结合多头自注意力机制,提出了一种预训练命名实体模型BERT-BiLSTM-self-Attention-CRF,通过对危险化学品的文本进行字符级别编码,得到基于上下文信息的字向量,增强了模型挖掘文本全局和局部特征的能力。实验结果表明,在自行构建的数据集上,本文模型优于其他传统模型,其$F1$值为94.57%。

关键词:命名实体识别;深度学习;危险化学品;预训练模型;自注意力机制

中图分类号:TP183　　**文献标识码**:A　　**文章编号**:1005-9164(2023)01-0043-09
DOI:10.13656/j.cnki.gxkx.20230308.005

人类知识学习是人工智能的研究方向之一,受人类解决问题行为方式的启发而形成的知识表示和推理,使智能系统能够准确地对获取的知识信息进行描述归类,并以此获得解决复杂问题的能力。谷歌公司在2012年提出知识图谱(Knowledge Graph,KG),这一结构化的人类知识形式引起了学术界和工业界的极大关注[1]。

随着硬件的发展,深度学习得到了长足的发展,相较于传统方法,深度学习在很多任务上都表现出巨大的优势[2]。主流的命名实体识别方法都是基于深度学习,将命名实体识别任务转化为序列标注任务,从而完成实体标签的预测[3]。但目前未有针对危险化学品的命名实体识别的相关研究,由于危险化学品的标签数据匮乏,现有的方法直接应用于危险化学品效果并不理想,实体的识别精度不高。针对这一问题,本文基于双向长短期记忆网络连接条件随机场(Bidirectional Long Short-Term Memory with Conditional Random Field,BiLSTM-CRF)模型[4],通过预训练的双向编码器表示(Bidirectional Encoder Representation from Transformers,BERT)[5]模型获取危险化学品领域的文本字符级别编码,得到基于上下文信息的字向量,并结合注意力机制,增强模型

收稿日期:2023-01-01　　修回日期:2023-01-07
*浙江省重点研发计划项目"危险化学品智慧监控及事故预防平台"(2020C03091)资助。
【第一作者简介】
陈观林(1978-),男,教授,主要从事人工智能、大数据分析研究,E-mail:chenguanlin@zucc.edu.cn。
【**通信作者】
【引用本文】
陈观林,程钊,邹凌,等.基于BERT的危险化学品命名实体识别模型[J].广西科学,2023,30(1):43-51.
CHEN G L,CHENG Z,ZOU L,et al.Named Entity Recognition Model of Hazardous Chemicals Based on BERT [J].Guangxi Sciences,2023,30(1):43-51.

挖掘文本的全局和局部特征的能力,以期促进危险化学品知识图谱的构建。

1 相关工作

命名实体识别任务的目的是从目标文本中识别出符合预定义语义类型的实体,是自然语言处理的一项基本任务,为下游的任务提供服务。目前,命名实体识别方法主要有以下 4 种。①基于规则的方法。该方法虽然不需要标注数据,但是需要依赖手工制作的规则,在规则的基础上制定相应的字典,提高识别的效果,如 LaSIE-Ⅱ[6]、NetOwl[7] 等。但是针对特定领域的规则,或者在领域词典不完整的情况下,这些系统通常具有较高的精确度和较低的召回率,并且这些系统无法转移到其他领域。②基于无监督学习的方法。该方法不需要手工去标记数据,典型的方法就是聚类,通过上下文的相似性,从聚类组中提取命名实体[8]。③基于特征的统计机器学习方法。该方法依赖于特征选择,应用监督学习可将命名实体识别任务转换成多分类或者序列标记任务,因此需要从文本中选取特征进行标注,并利用机器学习算法对特征进行训练,从而得到目标模型。常见的命名实体识别模型有马尔可夫模型[9]、支持向量机[10]、条件随机场(Conditional Random Field,CRF)[11] 等。④基于深度学习的方法,这也是现在最常用的方法。

与基于特征的学习方法比较,深度学习算法更能自动发现隐藏特征,并且效果也更好。1997 年提出的长短期记忆网络(Long Short-Term Memory,LSTM)模型可以选择遗忘前文中无用的信息,提高了命名实体识别的精度[12]。为解决 LSTM 模型只依赖之前的时序信息预测下一时刻输出,而无法考虑下文的状态等问题,Strubell 等[13] 将迭代碰撞卷积网络(Iterated Dilated Convolutional Neural Network,IDCNN)和 CRF 架构结合;Young 等[4] 将双向长短期记忆网络(Bidirectional Long Short-Term Memory,BiLSTM)和 CRF 架构结合;Devlin 等[5] 提出了采用 Transformer 编码和自注意力机制的 BERT 模型,通过对大规模语料进行训练,得到表征能力强的预训练字向量;谢腾等[14] 在 BiLSTM-CRF 架构的基础上,结合 BERT 模型,进一步提升了实体识别的精度。

基于深度学习算法的实体命名识别在诸多领域应用广泛,如医疗[15,16]、农业[17,18]、法律[19] 和生态[20] 等。但目前并没有学者针对危险化学品进行命名实体识别的相关研究,危险化学品领域也没有公开的、大规模的带标签的数据集。针对危险化学品的实体识别方法还存在着一些难点问题,如危险化学品领域内数据众多、存储格式各异、个体差异性较大。很多数据中信息实体间不在一个句子中,而且实体跟其他领域的实体类型重叠,与通用的命名实体也不一致,因此需要对概念实体重新定义。危险化学品缺乏足够多的标注数据,而人工标注的成本很高,需要花费大量的时间和精力,并且需要专业人员辅助标注,实体标注难度大。因此,本文针对危险化学品的命名实体识别问题,在 BiLSTM-CRF 的基础上,提出了 BERT-BiLSTM-self-Attention-CRF 模型。

2 BERT-BiLSTM-self-Attention-CRF 模型

本文提出的模型基于 BiLSTM-CRF,将 BiLSTM 的输入由 Word2Vec 预训练词向量替换成 BERT 预训练词向量,以此获取信息量更为丰富的词向量,同时在 BiLSTM 层后面接上自注意力机制层,从而更深层次地挖掘字符间的语义信息。

首先,向模型内输入字符进入 BERT 预训练模型,获取融合了字嵌入、分段嵌入和位置嵌入的编码字向量;其次,将融合语义的字向量作为 BiLSTM 网络的输入,通过 BiLSTM 可以让模型学习通过之前和之后的时序信息,并对下一时刻输出进行预测;再次,将 BiLSTM 网络挖掘到的全局特征,即 t 时刻的隐藏状态 h_t 作为输出,再通过自注意力机制层获取输出特征向量的相互影响关系,补充 BiLSTM 层输出向量的局部特征;最后,对 BiLSTM 层输出的标签预测变量,再通过 CRF 层,得出标签之间相互影响后出现的规律,如在 I-SUBJECT 后面不会接 B-SUBJECT 等,从而提高预测标签的合理性,使模型能够获得最佳的输出标签序列。模型结构如图 1 所示。

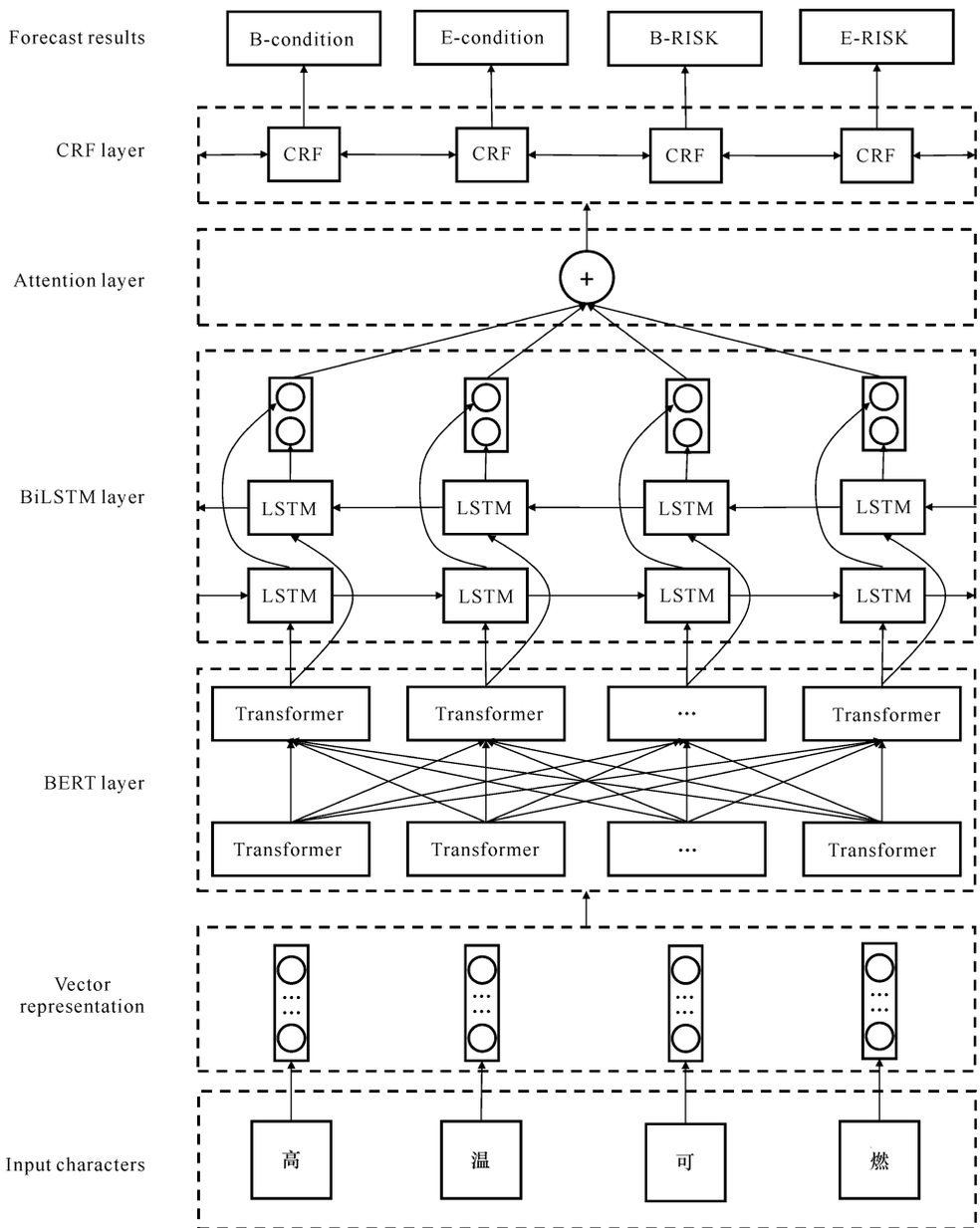

图 1　BERT-BiLSTM-self-Attention-CRF 模型结构

Fig. 1　Model structure of BERT-BiLSTM-self-Attention-CRF

2.1　基于 BERT 的字向量表示

在自然语言处理任务中,常用的词嵌入模型有 Google 提出的 Word2Vec[21] 和斯坦福大学提出的 Glove[22]。Word2Vec 通过词的上下文得到词的关联词向量;Glove 利用共现矩阵,同时考虑到局部信息和整体信息,丰富了词向量的语义信息。但是 Word2Vec 和 Glove 在单词长依赖场景表现均不是很好。而 BERT 的网络架构则使用 Vaswani 等[23]提出的多层 Transformer 结构,其最大的特点就是摒弃了传统的循环神经网络(Recurrent Neural Network,RNN)和卷积神经网络(Convolutional Neural Network,CNN),通过自注意力机制让任意两个位置的单词距离转换成 1,有效地解决了长依赖的问题。BERT 的输入向量嵌入是由字符向量嵌入、分段嵌入和位置嵌入的和组成,具体如图 2 所示。

BERT 预训练模型在进行自注意力计算时需要定义 3 个向量:Query 向量、Key 向量和 Value 向量。这 3 个向量是将词向量与训练后的 3 个权重矩阵 W_q、W_k 和 W_v 相乘得到自注意力层的计算公式,具体如下:

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_K}}\right)V, \quad (1)$$

式中,Q、K、V 分别为 Query、Key、Value 向量组合的矩阵;d 为输入向量即 Query 向量的维度,除以 $\sqrt{d_K}$ 可以有效控制梯度消失的问题,在训练过程中使模型梯度稳定下降。同时通过 softmax 函数将分数归一化,使输出的字向量可以充分学习到该字与其他字间的相互联系,从而丰富字语义表达。

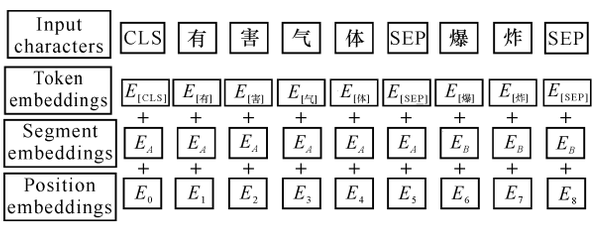

图 2 BERT 模型输入词向量

Fig. 2 BERT model input word vector

2.2 BiLSTM 层

随着输入序列长度的增加,传统的 RNN 会出现梯度爆炸和梯度消失的问题[24]。LSTM 在 RNN 的基础上进行改进,通过增加遗忘门、输入门和输出门 3 个门控单元,很好地解决了梯度消失和梯度爆炸的问题,加快了模型的收敛速度。LSTM 单元结构如图 3 所示。

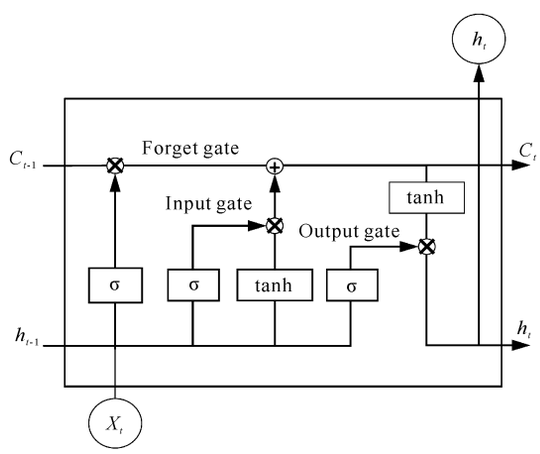

图 3 LSTM 单元结构

Fig. 3 LSTM unit structure

LSTM 单元结构的遗忘门、输入门和输出门的具体公式如下:

$$f_t=\sigma(W_f[h_{t-1},x_t]+b_f), \quad (2)$$

$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i), \quad (3)$$

$$o_t=\sigma(W_o[h_{t-1},x_t]+b_o), \quad (4)$$

式中,f_t、i_t、o_t 分别表示遗忘门、输入门和输出门;W_f、W_i、W_o 分别表示对应的权重矩阵;b_f、b_i、b_o 分别表示对应的偏置向量;σ 表示 sigmoid 激活函数;x_t 表示 t_n 时刻经过编码的输入字符向量;h_{t-1} 表示 t_{n-1} 时刻的隐藏层状态。

C_t 表示 t_n 时刻的细胞状态,更新公式如下:

$$C_t=f_t\odot C_{t-1}+i_t\odot \tanh(W_c[h_{t-1},x_t]+b_c), \quad (5)$$

式中,$\odot$ 表示哈达玛积;tanh 为双曲正切激活函数;W_c 和 b_c 分别表示更新状态的权重矩阵和偏置向量。

隐藏层状态的更新公式如下:

$$h_t=o_t\odot \tanh(C_t)。 \quad (6)$$

在命名实体识别任务中,当前的实体标签会同时受到上下文信息的影响。由于 LSTM 的结构特性,前向的 LSTM 只能考虑文本中的历史信息。因此本文采用 BiLSTM 模型,利用双向的 LSTM,同时学习文本的上下文信息,并对输出向量进行拼接,从而使字输出向量能够同时兼备上下文信息,克服了单向 LSTM 语义信息不完整的缺点,使实验结果更加准确。

2.3 多头自注意力机制层

在命名实体识别任务中,字符标签在很大程度上受到上下文某些词的影响,从而导致相同的字符在不同的语境下标签可能不同。由于预训练模型 BERT 是一个基于普通文本训练而得到的模型,因此通过 BERT 得到的词嵌入不能很好地表达危险化学品风险信息领域的语义。在获取上下文语义信息时,BiLSTM 网络更容易获取上下文的局部语义信息,但是却无法较好地表达输入文本序列的全局语义信息。也就是说,BiLSTM 网络不能很好地表达句子中每个字符对当前时间输出的重要性。因此,本研究将多头自注意机制层作为 BiLSTM 模块的一个附加模块,以此增强该模型挖掘全局信息和句子相关性的能力,使该模型能更好地应用于危险化学品风险信息领域。

在多头自注意力机制层中,Query 向量、Key 向量和 Value 向量分别使用不同的向量矩阵进行 h 次独立的线性映射,然后输入到 h 个并行头中执行自注意力操作。这样,每个并行头都可以获取输入的文本序列中各个字符在不同的表现空间中独有的特征语义信息。将 h 个并行头上的计算结果合并,并进行一个线性映射,得到最终的输出,具体的函数公式如下:

$$\mathrm{head}_i=\mathrm{Attention}(QW_i^Q,KW_i^K,VW_i^V), \quad (7)$$

$$\mathrm{Multihead}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)W^O, \quad (8)$$

式中,W_i^Q、W_i^K、W_i^V、W^O 表示线性变换所用到的权重

矩阵，$head_i$ 表示多头自注意力模块中的第 i 个头，Concat 表示拼接向量操作。

2.4 CRF 层

BiLSTM 层和多头自注意力机制层虽然可以学习上下文间的局部和全局特征信息，并输出对字的最大概率值的标签，但是不能学习各个标签间的关系，导致输出的连续标签不符合逻辑，存在同类型标签顺序错乱或者不同标签错误搭配的问题，如在 I-SUBJECT 后面接 B-SUBJECT 等。为了充分学习相邻标签间的依赖关系，本文在模型的最后采用 CRF 对多头自注意力机制层输出的特征信息进行解码，从而得到文本的标签序列。

在线性链条件随机场中，特征函数主要分为两大类，一类是定义在节点 x 上的状态特征函数，这类特征函数只与当前节点有关；另一类是定义在节点 y 上下文的转移特征函数，这类特征函数只与当前节点和上一节点有关。对于给定的输入序列 $X=x_1,x_2,\cdots,x_n$，可得到输出标签序列 $Y=y_1,y_2,\cdots,y_n$。

标签序列的得分函数可表示如下：

$$\text{score}(X,y)=\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i), \tag{9}$$

式中，t_k 表示局部特征函数；s_l 表示节点特征函数；λ_k、μ_l 分别是 t_k、s_l 的权重系数；k、l 分别代表转移特征函数和状态特征函数的个数。

对于给定的输入序列 X，可得到所有可能标签序列的分数，归一化公式如下：

$$p(y\mid X)=\frac{\exp(\text{score}(X,y))}{\sum_{\tilde{y}\in Y_X}\exp(\text{score}(X,\tilde{y}))}, \tag{10}$$

式中，$\text{score}(X,\tilde{y})$ 表示预测序列 $\tilde{y}$ 的评分函数得分，Y_X 表示所有可能的标注序列。

利用维特比(Viterbi)算法得到的最佳预测标签序列如下：

$$y^*=\text{argmax}(\text{score}(X,y))。 \tag{11}$$

3 结果与分析

3.1 数据集及标注体系

由于危险化学品文本样式多且格式不同，故本文以《危险化学品目录(2015 版)》[25]中记录的 2 828 种危险化学品为对象，爬取这些危险化学品对应的化学品安全技术说明书(Material Safety Data Sheets，MSDS)，经过数据清洗和预处理，过滤图片、重复信息等无用信息，并从中选取以句子为单位的危险化学品领域语料库。根据爬取数据的领域特性进行概念抽取，定义危险化学品领域的实体类别，具体如表 1 所示。

表 1 危险化学品领域实体类别

Table 1 Entity categories in the field of hazardous chemicals

实体类别 Entity category	类别描述 Category description	示例 Example
SUBJECT	Risk information initiator	Powder
CONTIDION	Occurrence conditions of risk information	High temperature
RISK	Risk information	Explode
RESULT	Risk product or risk result	Toxic gas

本文中的数据采用 BIEO (Begin，Inside，End，Outside)标注策略，考虑到采用的数据具有结构性，若进行正常的实体标注，则数据可分的种类较少，而且在抽取的语句中，多包含实体间的关系。故将实体间关系看作特殊的实体，定义为“条件”实体，通过实体识别的形式识别出“主体”“条件”“风险”“产物或结果”实体，并对这 4 种实体进行排列组合形成三元组，将形成三元组的步骤由先进行实体识别再进行关系抽取，改为先进行实体识别再进行实体排列组合[26]。这样能够充分利用语料语句信息，同时顾及实体和关系的特征，形成三元组阶段，直接对实体进行排列组合，相比于使用关系抽取构建的三元组速度更快。实体标签的定义如表 2 所示，其中 B 表示标注实体的开始部分，I 表示标注实体的中间部分，E 表示标注实体的结束部分，O 表示与实体无关的信息。

YEDDA[27]作为一种轻量级、高效、全面的文本跨度标注开源工具，从协同用户标注到管理员评估与分析，为文本范围标注提供了一个系统的解决方案。基于 YEDDA 设计出符合本研究的辅助实体标注平台，通过人工标注，并按照 9∶1 的比例将其划分为训练集和测试集。由于收集的数据比较少，首先，将训练集中同类型实体相互替换[28]获取新的语料以扩充数据集，替换比例为 20%；其次，利用现有数据集训练出模型后，再利用训练得到的模型对未标注的语句进行标注；再次，通过人工复查的方式确保模型标注后的语句的正确性；最后，将这些数据重新放入数据集中从而获得最终数据集。整体流程如图 4 所示，危险化学品领域数据集中有主体标签实体 350 个，条件标签实体 7 033 个，风险标签实体 4 156 个，产物或结果标签实体 1 674 个。数据集划分为训练集和测

试集,训练集含有 3 216 条句子,测试集含有 360 条句子。

表 2　危险化学品领域实体类别定义

Table 2　Entity category definitions in the field of hazardous chemicals

序号 No.	实体标签 Entity label	标签含义 Meaning of label
1	B-SUBJECT	First character of subject
2	I-SUBJECT	Non first and last character of subject
3	E-SUBJECT	Last character of subject
4	B-CONDITION	First character of condition
5	I-CONDITION	Non first and last character of condition
6	E-CONDITION	Last character of condition
7	B-RISK	First character of risk
8	I-RISK	Non first and last character of risk
9	E-RISK	Last character of risk
10	B-RESULT	First character of result
11	I-RESULT	Non first and last character of result
12	E-RESULT	Last character of result
13	O	Non entity character

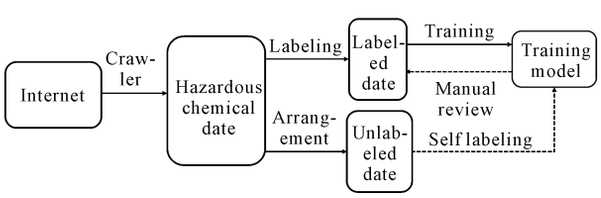

图 4　数据集构建流程

Fig. 4　Dataset construction process

3.2　评价指标

通常使用精确率和召回率来衡量命名实体识别模型的效果,精确率是指模型预测正确的实体个数占所有待识别的实体个数的比例,召回率是指模型预测正确的实体个数占标注的实体个数的比例。不同的场景下看重的指标有所不同,为了综合评价模型的性能,故将精确率和召回率同等看待,计算精确率和召回率的加权几何平均值即 $F1$ 值。本实验以 $F1$ 值作为命名实体识别的评价指标,各指标的计算公式如下:

$$\text{精确率}=\frac{\text{预测正确的实体个数}}{\text{总的实体个数}},\tag{12}$$

$$\text{召回率}=\frac{\text{预测正确的实体个数}}{\text{标注的实体总个数}},\tag{13}$$

$$F1=\frac{2\times\text{精确率}\times\text{召回率}}{\text{精确率}+\text{召回率}}。\tag{14}$$

3.3　实验结果与分析

为进一步验证 BERT-BiLSTM-self-Attention-CRF 模型对每类实体的识别效果,在模型参数数量相等的情况下,与 IDCNN-CRF[11]、BiLSTM-CRF[4]、BERT-CRF[5]、BERT-BiLSTM-CRF[14]模型进行比较,对比结果如表 3 所示。BERT-BiLSTM-self-Attention-CRF 模型在识别 CONDITION 实体、RISK 实体以及平均表现均优于其他模型。值得注意的是,在实体数较少的 SUBJECT 实体类别的识别中,IDCNN-CRF 模型表现优于其他模型,这主要是由于在样本少的情况下,IDCNN 模型对语料信息的抽取能力优于 BiLSTM 模型。而 BERT-BiLSTM-CRF 模型在 RESULT 实体的识别上效果优于其他模型,甚至优于加入自注意力机制层的模型,可能是由于语料中 RESULT 实体大多属于长单词,经过自注意力机制层,长单词中的每个字符所包含的注意力信息相互叠加,对识别实体类别造成了一定的影响,而 CONDITION 实体、RISK 实体大多属于短单词,因此所有加入自注意力机制层的 BERT 模型表现更好。

3.4　危险化学品知识图谱构建

化学品安全技术说明书是结构化的文档,在构建危险化学品知识图谱时,参考刘宝等[29]筛选的危险化学品的属性进行初步构建,同时将训练好的模型对危险化学品对应的化学品安全技术说明书中挑选出的长文字句子进行实体识别。由于关系也被定义成实体类型,故可以对识别结构做拼接处理,组合成三元组形式的数据并存储,形成危险化学品领域知识图谱。

图 5 展示了知识图谱单节点的结构,不同颜色的圆代表不同的实体,连接两圆的直线代表两实体间的关系。其中,淡绿色表示危险化学品实体,淡粉红表示危害,蓝色表示预防措施,黄色表示事故响应措施,粉色表示存放要求,深绿色表示废弃要求,棕色表示特殊危险性。

表 3　模型实体类别识别结果对比(%)

Table 3　Comparison of model entity category identification results (%)

实体类别 Entity category	模型 Model	精确率 Accuracy	召回率 Recall	*F*1
SUBJECT	IDCNN-CRF	69.23	75.00	72.00
	BiLSTM-CRF	50.00	50.00	50.00
	BERT-CRF	60.00	50.00	54.55
	BERT-BiLSTM-CRF	75.00	50.00	60.00
	BERT-BiLSTM-self-Attention-CRF	**75.00**	**50.00**	**60.00**
CONDITION	IDCNN-CRF	89.94	88.82	89.38
	BiLSTM-CRF	86.87	92.47	89.58
	BERT-CRF	91.62	94.09	92.84
	BERT-BiLSTM-CRF	91.15	94.09	92.59
	BERT-BiLSTM-self-Attention-CRF	**92.67**	**95.16**	**93.90**
RISK	IDCNN-CRF	92.86	90.70	91.76
	BiLSTM-CRF	95.41	96.30	95.85
	BERT-CRF	96.26	95.37	95.81
	BERT-BiLSTM-CRF	96.26	95.37	95.81
	BERT-BiLSTM-self-Attention-CRF	**95.41**	**96.30**	**95.85**
RESULT	IDCNN-CRF	68.42	86.67	76.47
	BiLSTM-CRF	91.67	91.67	91.67
	BERT-CRF	95.92	97.92	96.91
	BERT-BiLSTM-CRF	97.96	100.00	98.97
	BERT-BiLSTM-self-Attention-CRF	**97.92**	**97.92**	**97.92**
All	IDCNN-CRF	88.36	88.69	88.52
	BiLSTM-CRF	89.47	92.82	91.11
	BERT-CRF	93.18	94.25	93.71
	BERT-BiLSTM-CRF	93.47	94.54	94.00
	BERT-BiLSTM-self-Attention-CRF	**94.03**	**95.11**	**94.57**

Note:results of this article are displayed in bold font

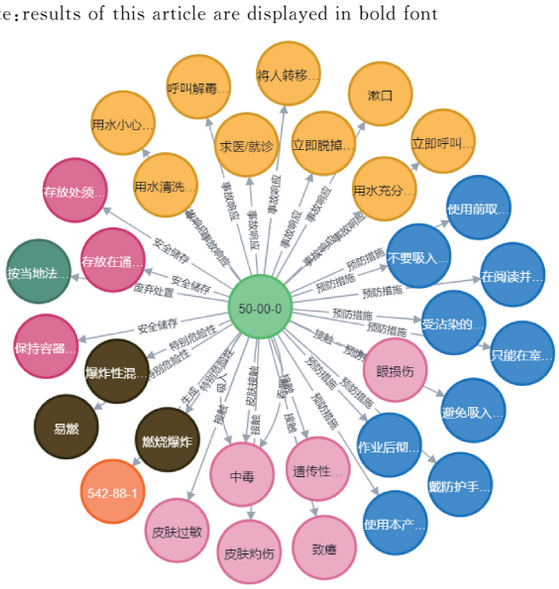

图 5　甲醛溶液节点信息

Fig. 5　Node information of formaldehyde solution

4　结论

基于危险化学品数据集,本文提出了一种 BERT-BiLSTM-self-Attention-CRF 的命名实体识别模型,在 BiLSTM-CRF 模型的基础上引入 BERT 预训练语言模型,以丰富初始字向量的语义特征,利用预训练模型获取初始化的词向量,克服了危险化学品领域语料匮乏的问题,并采用自注意力机制,捕捉字向量之间的内部相关性,让模型能够高度关注关联性高的重要信息。实验结果表明,BERT-BiLSTM-self-Attention-CRF 模型能够很好地识别危险化学品的实体,精确率为 94.03%,召回率为 95.11%,*F*1 值为 94.57%,满足研究需求。在未来的研究中,仍须进一步扩大和完善危险化学品数据集,并在此基础上进行事件抽取,构建面向危险化学品的知识图谱。

参考文献

[1] 马忠贵,倪润宇,余开航. 知识图谱的最新进展、关键技术和挑战[J]. 工程科学学报,2020,42(10):1254-1266.

[2] JI S X,PAN S R,CAMBRIA E,et al. A survey on knowledge graphs:representation,acquisition,and applications [J]. IEEE Transactions on Neural Networks and Learning Systems,2021,33(2):494-514.

[3] LI J,SUN A X,HAN J L,et al. A survey on deep learning for named entity recognition [J]. IEEE Transactions on Knowledge and Data Engineering,2020,34(1):50-70.

[4] YOUNG T,HAZARIKA D,PORIA S,et al. Recent trends in deep learning based natural language processing [J]. IEEE Computational Intelligence Magazine,2018,13(3):55-75.

[5] DEVLIN J,CHANG M W,LEE K,et al. BERT:pre-training of deep bidirectional transformers for language understanding [C]//Conference on the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. [S. l.]:NAACL,2019:4171-4186.

[6] GÜNGÖR O,ÜSKÜDARLI S,GÜNGÖR T. Improving named entity recognition by jointly learning to disambiguate morphological tags [C]//Proceedings of the 27th International Conference on Computational Linguistics. New York:ACM,2018:2082-2092.

[7] CHEN Y P,DING Z H,ZHENG Q H,et al. A history and theory of textual event detection and recognition [J]. IEEE Access,2020,8:201371-201392.

[8] YADAV V,BETHARD S. A survey on recent advances in named entity recognition from deep learning models [C]//Proceedings of the 27th International Conference on Computational Linguistics. New York:ACM,2018:2145-2158.

[9] SONG H J,JO B C,PARK C Y,et al. Comparison of named entity recognition methodologies in biomedical documents [J]. Biomedical Engineering Online,2018,17(Suppl 2):158.

[10] LEÓN F S,LEDESMA A G. Annotating and normalizing biomedical NEs with limited knowledge [C]//Proceedings of the 5th Workshop on BioNLP Shared Tasks. Stroudsburg,PA:ACL,2019:62-71.

[11] ZHAO S D,LIU T,ZHAO S C,et al. A neural multi-task learning framework to jointly model medical named entity recognition and normalization [C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence. Menlo Park,CA:AAAI,2019:817-824.

[12] SHERSTINSKY A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network [J]. Physica D:Nonlinear Phenomena,2020,404:132306.

[13] STRUBELL E,VERGA P,BELANGER D,et al. Fast and accurate entity recognition with iterated dilated convolutions [C]//Conference on Empirical Methods in Natural Language Processing. Stroudsburg,PA:ACL,2017:2660-2670.

[14] 谢腾,杨俊安,刘辉. 基于 BERT-BiLSTM-CRF 模型的中文实体识别[J]. 计算机系统应用,2020,29(7):48-55.

[15] 巩敦卫,张永凯,郭一楠,等. 融合多特征嵌入与注意力机制的中文电子病历命名实体识别[J]. 工程科学学报,2021,43(9):1190-1196.

[16] 孙超,张文博. 中医古籍文本术语命名实体识别的研究进展与挑战[J]. 中华中医药杂志,2021,36(11):6843-6845.

[17] 李林,周晗,郭旭超,等. 基于多源信息融合的中文农作物病虫害命名实体识别[J]. 农业机械学报,2021,52(12):253-263.

[18] 赵鹏飞,赵春江,吴华瑞,等. 基于注意力机制的农业文本命名实体识别[J]. 农业机械学报,2021,52(1):185-192.

[19] 李春楠,王雷,孙媛媛,等. 基于 BERT 的盗窃罪法律文书命名实体识别方法[J]. 中文信息学报,2021,35(8):73-81.

[20] 蒋翔,马建霞,袁慧. 基于 BiLSTM-IDCNN-CRF 模型的生态治理技术领域命名实体识别[J]. 计算机应用与软件,2021,38(3):134-141.

[21] SARZYNSKA-WAWER J,WAWER A,PAWLAK A,et al. Detecting formal thought disorder by deep contextualized word representations [J]. Psychiatry Research,2021,304:114135.

[22] GU Y,TINN R,CHENG H,et al. Domain-specific language model pretraining for biomedical natural language processing [J]. ACM Transactions on Computing for Healthcare,2021,3(1):1-23.

[23] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need [C]//31st Conference on Neural Information Processing Systems. Cambridge,MA:MIT Press,2017:1-11.

[24] KHAN A,SOHAIL A,ZAHOORA U,et al. A survey of the recent architectures of deep convolutional neural networks [J]. Artificial Intelligence Review,2020,

53(8):5455-5516.
[25] 危险化学品目录(2015版)[EB/OL].(2015-02-27)[2022-08-25].http://www.chinasafety.gov.cn.
[26] 刘奔,姬东鸿.药物实体和药物相互关系的联合识别[J].计算机工程与设计,2017,38(5):1377-1381.
[27] YANG J,ZHANG Y,LI L W,et al. YEDDA:a lightweight collaborative text span annotation tool [C]// 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg,PA:ACL,2017:31-36.
[28] ZHANG X,ZHAO J B,LECUN Y. Character-level convolutional networks for text classification [C]// 29th Annual Conference on Neural Information Processing Systems. Cambridge,MA:MIT Press,2015:649-657.
[29] 刘宝,车礼东,黄红花,等.基于自然语言处理(NLP)技术建立化学品危险评估知识图谱的研究[J].计算机与应用化学,2018,35(7):605-610.

Named Entity Recognition Model of Hazardous Chemicals Based on BERT

CHEN Guanlin[1,2**],CHENG Zhao[1,2],ZOU ling[2],YANG Wujian[1],LI Tian[1]
(1. School of Computer and Computing Science,Hangzhou City University,Hangzhou,Zhejiang,310015,China;2. School of Computer Science and Artificial Intelligence,Changzhou University,Changzhou,Jiangsu,213164,China)

Abstract:Aiming at the problems of hazardous chemicals entity recognition and relationship recognition,based on the Bidirectional Long Short-Term Memory with Conditional Random Field (BiLSTM-CRF) model,the pre-training named entity model BERT-BiLSTM-self-Attention-CRF is proposed by introducing the Bidirectional Encoder Representation from Transformers (BERT) model in combination with the multi-head self-attention mechanism. By encoding the text of hazardous chemicals at the character level,the character vector based on context information is obtained,which enhances the ability of the model to mine the global and local features of the text. The experimental results show that the proposed model is superior to other traditional models on the self-built data set,and its $F1$ value is 94.57%.

Key words:named entity recognition;deep learning;hazardous chemicals;pre-training model;self-attention mechanism

责任编辑:唐淑芬