基于混合神经网络和注意力机制的生物医学事件触发词识别方法

任永功[1],林禹竹[1],唐玉洁[1],于 博[1],何馨宇[1,2,3*]
(1. 辽宁师范大学计算机与人工智能学院,辽宁大连 116081;2. 大连理工大学通信与工程博士后研究站,辽宁大连 116081;
3. 大连永佳电子技术有限公司博士后工作站,辽宁大连 116081)

摘 要: 生物医学事件作为生物医学文本挖掘的重要组成部分,在生物医学研究和疾病的预防中发挥着重要作用. 触发词识别是生物医学事件抽取的关键和前提步骤,旨在提取描述事件类型的关键词. 传统方法在特征提取过程中过分依赖自然语言处理工具,导致耗费人工成本. 另外,由于生物医学文献的特殊性—长文本语句多,导致长距离依赖问题比较明显. 为了解决这些问题,我们提出了一种混合结构,由残差卷积神经网络和双向长短期神经网络、混合神经网络和多头注意力机制组成. 该模型利用残差卷积神经网络提取单词级特征并利用双向长短期神经网络提取上下文语义信息. 此外,本文通过空间域滑动窗口将长句划分为等长短句,在不破坏上下文信息的前提下,避免了长距离依赖. 实验结果表明,本文提出的方法在生物医学事件抽取通用语料 MLEE(Multi-Level Event Extraction)上取得了较好的效果,*F* 值达到 81.15%.

关键词: 生物医学事件抽取;触发词识别;ReCNN-BiLSTM;空间域滑动窗口;MUH-Attention 机制;混合神经网络
基金项目: 国家自然科学基金(No.62006108,No.61976109);辽宁省"兴辽英才计划"项目(No.XLYC2006005);辽宁省普通高等教育本科教学改革研究项目(辽教通[2022] 166 号);辽宁省高等学校科学研究项目(No.LJKZ0963);辽宁师范大学本科教学改革研究与实践项目(No.LSJG202210);辽宁省科技厅重点研发项目(No.2022JH2/101300271)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)09-3206-11
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221361

A Biomedical Event Trigger Identification Method Based on Hybrid Neural Network and Attention Mechanism

REN Yong-gong[1], LIN Yu-zhu[1], TANG Yu-jie[1], YU Bo[1], HE Xin-yu[1,2,3*]
(1. *School of Computer and Information Technology*, *Liaoning Normal University*, *Dalian*, *Liaoning* 116081, *China*;
2. *Information and Communication Engineering Postdoctoral Research Station*, *Dalian University of Technology*, *Dalian*, *Liaoning*, 116081, *China*;
3. *Postdoctoral Workstation of Dalian*, *Yongjia Electronic Technology Co. , Ltd. , Dalian*, *Liaoning* 116081, *China*)

Abstract: Biomedical events, as an important part of biomedical text mining, play an important role in biomedical research and disease prevention. Trigger identification is the key and prerequisite step of biomedical event extraction, which aims to extract the key words describing event types. Traditional trigger identification methods rely too much on natural language processing tools in the process of feature extraction, consuming a lot of manual cost. In addition, due to the particularity of biomedical literature—there are many long text sentences, the problem of long-distance dependence is obvious. To solve these problems, we propose a hybrid structure, which is composed of residual convolution neural network and bidirectional long short term memory, hybrid neural network and multi head attention mechanism. The proposed model uses residual convolution neural network to extract vocabulary-level features and bidirectional long short term memory to obtain contextual semantic information. Furthermore, spatial domain sliding windows divide long sentences into equal-length short sentences without damaging context information, which can avoid long-distance dependency without destroying the context information. The experimental results show that our method outperforms the state-of-the-art methods on the commonly used multi-level event extraction (MLEE) corpus, achieving 81.15% *F*-score.

收稿日期:2022-12-01;修回日期:2023-03-14;责任编辑:梅志强
*通讯作者:何馨宇

Key words: biomedical event extraction; trigger detection; ReCNN-BiLSTM; spatial domain sliding window; MUH-Attention; mixed neural network

Foundation Item(s): National Science Foundation of China (No.62006108, No.61976109); Liaoning Revitalization Talents Program (No.XLYC2006005); Research Project of Liaoning General Higher Education (No.Liao Jiao Tong [2022] 166); Scientific Research Project of Liaoning Province (No.LJKZ0963); Research and Practice Project of Undergraduate Teaching Reform of Liaoning Normal University (No.LSJG202210); Key R&D Projects of Liaoning Provincial Department of Science and Technology (No.2022JH2/101300271)

1 引言

近年来,随着网络和信息技术的飞速发展,生物医学领域受到越来越多研究者的关注. 生物医学文献的数量呈指数级增长,使得相关研究人员很难从大量的生物学文献中快速获得有用的知识. 因此,生物医学文本挖掘技术应运而生. 生物医学文本挖掘又称生物医学自然语言处理 BioNLP(Biomedical Natural Language Processing),主要研究如何从大量生物医学文献中自动提取有用信息,供生物医学研究人员查询和研究. 生物医学领域信息抽取的最终目的是以结构化的形式呈现研究者感兴趣的非结构化信息,以提高研究效率. 信息抽取包括实体识别、关系抽取,事件抽取等. 生物医学事件抽取任务旨在自动地对生物医学文本中包含的事件进行详细的分析和抽取,对生物医学的研究领域具有重大意义. 生物医学事件触发词识别是生物事件抽取中基础且关键的步骤,其主要目的是从生物医学文献中自动地识别出相应的触发词,并正确判定其类型. 生物医学领域的触发词识别任务通常被形式化为多分类问题. 在实际操作中,我们采用"BIO"标记的方法来解决触发词包含多个单词的问题,该方法在关系抽取任务上取得了较好的效果[1]. 在"BIO"标记法中,B 表示触发词短语的第一个单词,I 表示触发词短语的中间单词,O 表示该单词不是触发词.

生物医学事件抽取主要研究生物分子间细粒度交互关系. 生物医学事件是指一个或多个生物医学实体的状态变化,包括基因表达、转录、分解代谢、磷酸化、定位、结合和调控等. 如 MLEE(Multi-Level Event Extraction)语料库所述,生物医学事件包含一个触发词和一个或多个要素. 事件触发词是用来表征生物事件发生的词或短语,通常是动词或动名词. 如图 1 所示,句子中有 3 个生物医学事件:第 1 个事件是基因表达事件 *T*1,包括一个触发词"production"和一个主题类型的要素"PROTE-10";第 2 个事件是正调控事件 *T*2,包括触发词"induction"、主题类型要素 *T*1 和原因要素"Cdc41";最后一个事件是负调控事件 *T*3,包括触发词"prevented"和主题类型要素 *T*2. 事件 *T*1 是简单事件,*T*2 和 *T*3 是复杂事件,他们的要素嵌套了其他事件. 3 个事件的结构如下:

Event *T*1 (Type: Gene_expression, Trigger: production, Theme: PROTE-10);

Event *T*2 (Type: Positive_regulation, Trigger: induction, Theme: Event *T*1, Cause: Cdc41);

Event *T*3 (Type: Negative_regulation, Trigger: prevented, Theme: Event *T*2).

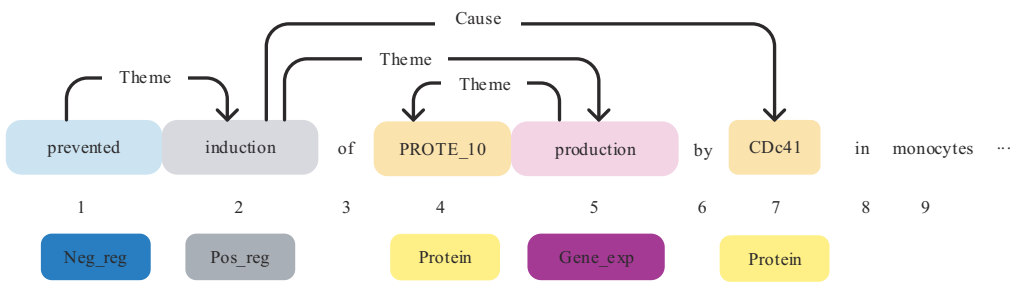

图 1　生物医学事件示例

触发词是一个事件或一个动作的发起词语,在句子中充当谓语的角色,通常为动词或者动名词. 生物医学触发词识别的性能直接影响着生物医学事件抽取的整体性能. 文章中的触发词一般具有以下三大特点:(1)可读性. 关键词本身应该是有意义的词或者短语. 例如,在"我们的结果表明,α-MSH 在一定程度上影响对细胞间连接活性的激活起调节作用"这一句话中,"α-MSH"不是触发词,"调节"是触发词;(2)上下文语义的相关性. 触发词必须与所在句子或者所在文段语句存在联系;(3)高度覆盖性:触发词是一个简单事件的发起词或者是嵌套事件所包含的一个子事件的发起词,事件的发起词要在事件中起到重要的覆盖性的作用,

且能够起到覆盖整个事件的作用. 触发词识别方法主要分为四类:基于词典的方法、基于规则的方法、基于统计机器学习的方法和基于深度学习的方法. 近年来,随着深度学习的不断发展,各种各样的神经网络应运而生. 同时,利用深度学习对实验进行预训练的方法也越来越多. 这些方法具有学习能力强、覆盖范围广、适应性强、可移植性好等优点,使得原始语义词嵌入的预训练更具包容性,提高了触发词识别任务的准确性. Bengio 等人[2]首先使用神经网络建立了语言模型. Mikolov 等人[3,4]在 2013 年提出了基于 Log-Bilinera 模型的 Word2vec 技术[5]. 本文利用双向长短期神经网络 BiLSTM(Bidirectional Long Short Term Memory)学习卷积神经网络 CNN(Convolution Neural Networks)合并后的特征,用以获取句子的特征表示.

现有的触发词识别方法主要分为基于传统机器学习的方法和基于深度学习的方法. 传统的触发词识别方法耗费大量的人工成本,Pyysalo 等人[6]提出了一种基于支持向量机 SVM(Support Vector Machines)的方法,将人工设计的显著特征输入到 SVM 分类器中进行事件触发词识别. He 等人[7]提出了一种基于 SVM 和 PA(Passive Aggressive alogrithm)算法的两阶段触发词识别方法. Zhou 等人[8]提出了一种新的 SVM 事件触发词识别框架,并结合了领域知识提升识别性能. 这类方法通过自然语言工具包来进行特征分类,不仅时间长还消耗大量的人工资源,模型泛化性能较差. 因此,基于深度学习的触发词识别方法相继出现. Nie 等人[9]首先提出了一种人工神经网络预测模型进行触发词识别. Rahul 等人[10]设计了一种利用递归神经网络 RNN(Recurrent Neural Network)对句子中的高级特征进行提取的触发词识别方法. Wei 等人[11]提出了一种多层剩余双向长短时记忆(BiLSTM)结构. Chen 等人[12]构建了动态多池化卷积神经网络(DMCNN). 为了充分吸收各种神经网络在触发词识别任务中体现的优势,一些研究者提出了混合神经网络. Diao 等人[13]设计了基于 LSTM(Long Short Term Memory)和 SVM 的混合神经网络结构. Shen 等人[14]提出了一种端到端卷积神经网络和极限学习机(CHNN-ELM)框架检测生物医学事件触发词. 本文也综合应用了 ReCNN(Residual Convolutional Neural Network)局部特征的强学习能力和 BiLSTM 和对于前后文特征的提取优势构建了触发词识别混合神经网络.

以上方法各有优点,然而由于生物医学文献中存在的长文本导致的长距离依赖问题并未得到很好的解决,为此,本文引入了空间域滑动窗口. 该方法在保证语义信息不丢失的前提下将长句划分为若干等长短句,有效解决了长距离依赖问题. 此外,触发词的位置信息以及上下文单词的语义信息对触发词类型的判断具有重要意义. 例如,在下面这句话中,Our results indicate that alpha-MSH exerts modulatory effects {{Positive_regulation}} on the activation {{Positive_regulation}} of intercellular junction activities effects {{Regulation}} to a certain extent…,有 3 个触发词:第 1 个“effects”、“activation”和第 2 个“effects”. 由于单词“activation”的语义影响,第 1 个“effects”被判定为 Positive_regulation 类型. 但是,由于从“activation”到第 2 个“effects”的距离较远,语义信息的影响不大,因此第 2 个“effects”被归类为 Regulation 类型. 由此可见,单词的位置及语义信息直接影响着触发词类型的识别准确度. 因此,本文设计了多头注意力 MUH-Attention(Multi Head Attention)机制获得每个单词的位置信息特征,进一步提升触发词识别性能. 综上,本文提出了一种新颖且有效的混合神经网络,它由空间域滑动窗口、MUH-Attention、ReCNN 和 BiLSTM 四部分组成. 该模型利用空间域滑动窗口将长句划分为等长短句,解决了长距离依赖问题;利用 ReCNN 提取词级特征,引入 BiLSTM 训练上下文信息,通过 MUH-Attention 机制获得每个词的位置信息特征. 上述混合神经网络充分利用了 ReCNN 局部特征的强学习能力,利用 BiLSTM 神经网络获取前向和后向特征,提取句子的特征表示,进而提高了触发词识别的性能. 本文的主要贡献如下:

(1)提出了一种基于 ReCNN 和 BiLSTM 的混合神经网络. 该网络具有两个主要优点:一方面,ReCNN 利用自身的特点提高了单词级特征的识别准确率,加快了模型的收敛速度和损失值的下降速度. 另一方面,BiLSTM 在解决长序列训练过程中的梯度消失等问题有着良好表现,同时可以更好的获取双向语义信息.

(2)本文在上述混合神经网络的基础上引入了空间域滑动窗口,将单词分批输入到神经网络中. 空间域滑动窗口通过将生物医学文献中的长句划分为若干相同长度的短句,解决了长距离依赖问题,防止上下文语义信息消失.

(3)引入了 MUH-Attention 机制. 该机制计算句子中所有单词的注意力,并增强单词的位置和语义信息. MUH-Attention 机制为每个单词提供更为准确的位置信息,进而提高了生物医学事件触发词识别的准确性.

2 相关工作

在大多数事件抽取方法中,触发词识别被公认为是一项多分类任务. 在生物医学事件抽取通用数据集 MLEE 上,Pyysalo 等人[6]提出了一种基于支持向量机 SVM 的方法,将人工设计的显著特征(如上下文特征)输入到一对一的 SVM 分类器中进行事件触发词识别.

He等人[7]提出了一种基于SVM和PA算法的两阶段触发词识别方法. 在第一阶段,对事件中的候选词进行二分类,判断其是否为触发词. 在第二阶段,通过多分类确定具体的触发词类别. Zhou等人[8]提出了一种新的事件触发词识别框架,该框架从Medline建立的大型文本语料库中学习生物医学领域知识,并通过神经网络建模将其嵌入到单词特征中.

上述基于传统机器学习的触发词识别方法需要有专业领域知识的技术人员设计大量的特征,人工成本较高,且系统泛化性能相对较差. 近年来,基于深度学习的方法受到了广大研究者的关注. Nie等人[9]首先提出了一个基于词向量辅助的人工神经网络预测模型(EANNP),在MLEE上取得很好的效果. Rahul等人[10]提出了一种利用递归神经网络RNN对句子中的高级特征进行提取的方法,该方法利用RNN的隐藏状态表示,以单词嵌入和实体类型作为特征,进而避免了使用各种NLP工具箱生成复杂的人工特征. Wang等人[15]利用神经网络结构在原有特征依赖的基础上学习更好的词嵌入特征表示,利用SoftMax分类器对触发词进行识别和分类. Li等人[16]提出了一种上下文标记敏感门控网络,该网络根据候选触发词动态调整,自动捕获上下文标记线索,并引入依赖词嵌入表达语义信息和注意力机制,进而获得更集中的表示. Chen等人[12]提出了一种动态多池化卷积神经网络(DMCNN),他们利用卷积神经网络CNN捕获句子级线索,根据事件的触发词和要素的变化动态设置多池化层的参数,获得了较好的实验性能. Wei等人[11]引入语言模型动态计算上下文的单词表示,提出了一种多层剩余双向长短时记忆BiLSTM结构,有效地解决了生物医学语料库中标记的歧义问题. 为了充分吸收各种神经网络在触发词识别任务中体现的优势,一些研究者提出了混合神经网络. 张等人[17]设计了一种融合无监督和有监督两种学习方式的混合监督深度信念网络(DL-ERM),获得文本特征数据的深度特征,以识别事件. Diao等人[13]提出了一种基于SVM和BiLSTM的混合结构(FSBN)来识别生物医学事件触发词. 该混合结构通过细粒度表示抽取更高层次和更准确的特征,并利用支持向量机对生物医学事件触发词的小数据集结果进行分类. Shen等人[14]提出了一种端到端卷积神经网络和极限学习机(CHNN-ELM)框架检测生物医学事件触发词. 利用(CHNN)特有的四个不同维度的神经网络层,有效地选择语义特征;利用ELM(Extreme Learning Machine)的可扩展性和泛化性能进行生物医学事件触发词识别,在MLEE语料上取得了不错的效果. 受以上工作的启发,本文也综合应用了ReCNN和BiLSTM的优势构建了混合神经网络. 通过BiLSTM获取双向语义信息;通过ReCNN获取局部特征,加快模型的收敛速度和损失值的下降速度.

以上方法均有各自的优势,尤其在研究方法上都有着重要的借鉴意义. 然而生物医学文本中普遍存在的远程依赖问题并没有得到很好的解决,这可能导致触发词识别错误,为此本文引入了空间域滑动窗口,在不破坏语义间上下文信息和位置信息的前提下,将生物医学领域中的长句划分为多个等长短句,解决由于长距离依赖引起的重要信息丢失问题. 使用混合神经网络ReCNN和BiLSTM解决了传统方法在特征提取中过于依赖自然语言处理工具的问题,减少了人工成本,同时综合利用了ReCNN和BiLSTM神经网络的优势提高了单词特征的识别准确率. 此外,由于触发词的位置信息对触发词类型的判断具有重要意义,本文通过构建MUH-Attention机制来提供更为准确的位置信息,进一步提高了生物医学事件触发词的识别性能.

3　方法

图2为本文提出的触发词识别模型. 首先通过BioBERT将从PubMed上下载的5.7 GB背景语料与原始语料库MLEE合并训练成词向量. 然后,通过空间域滑动窗口对训练降维后的词向量进行分批处理,作为混合神经网络ReCNN的输入. 随后,将ReCNN处理后的单词级词向量加入MUH-Attention机制. 同时,采用MUH-Attention处理后的单词级词向量作为BiLSTM的输入,利用逆高阶窗(矩阵)获得前向累积信息和预测相关信息,使上下文信息更加紧凑,得到隐含层表示. 最后利用全连接层和SoftMax对触发词进行分类.

3.1　BioBERT训练词嵌入

近年来,词嵌入技术应用广泛. 通过词嵌入,可以从大量未标记的数据中获取丰富的语义特征信息. 常用的词嵌入包括Collobert和Weston嵌入[18]、HLBL嵌入[19]和Word2Vec[5,20]等. 然而,由于生物医学文本的特殊性,目前常用的词嵌入训练方法对生物医学词汇的挖掘效果并不理想. BioBERT是BERT[21]针对生物医学语料库的预训练改进工具,有助于理解复杂的生物医学文献. BioBERT在生物医学命名实体识别(F值提高0.62%)、生物医学关系抽取(F值提高2.80%)和生物医学文本挖掘三个具有代表性的任务中均取得了良好的效果[22]. 因此,本文使用BioBERT训练词嵌入$\boldsymbol{x}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_n]$,通过单词嵌入来丰富单词嵌入的上下文语义信息$\boldsymbol{t}=[\boldsymbol{t}_1,\boldsymbol{t}_2,\boldsymbol{t}_3,\cdots,\boldsymbol{t}_n]$,其中$n$是单词嵌入的数量,最后生成具有丰富语义信息的单词嵌入$\boldsymbol{x}_s=\boldsymbol{x}\oplus\boldsymbol{t}$,$\oplus$表示向量的点乘.

3.2　ReCNN

ReCNN可以提取句子中每个单词的局部特征,进而提高句子特征的准确性;同时可以使模型快速收敛,

提高训练速度. 在ReCNN中,卷积相当于使用卷积核在输入矩阵上滑动积和的过程. 本文将残差卷积神经网络中的Kernel_size(卷积核大小)、In_channel(输入通道)以及Out_channel(输出通道)的参数进行适应性调整得到了ReCNN,用来提取句子中每个词的局部特征;然后用特征映射的形式表示每个单词的词嵌入. 本文通过加入残差神经网络实现跨层的身份映射,提取单词级特征,最后以特征映射的形式表达. 与CNN相比,ReCNN实现了跨层的身份映射,大大降低了训练难度. 此外,通过使用ReCNN减少过拟合,使得特征提取更加准确.

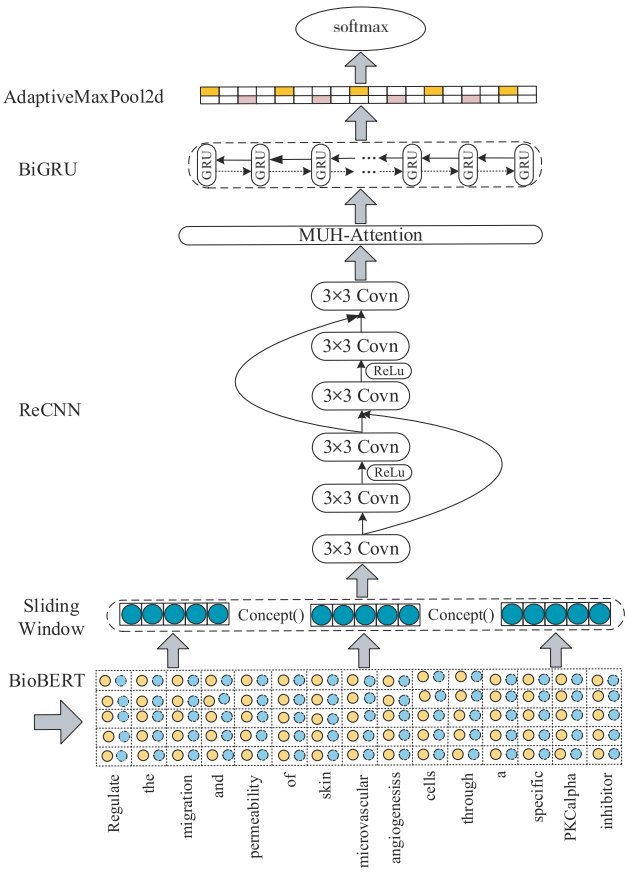

图2 触发词识别模型

(1)输入层. 通过空间域滑动窗口将生物医学文本句子分割成短句,作为神经网络的输入. 本文将所用的空间域滑动窗口的大小设置为5. 句子中每个单词的词嵌入表示为$\boldsymbol{x}_i \in \mathbf{R}^{n\times d}$,其中$n$为字数,$d$为嵌入维度,本文定义为768维.

(2)卷积层. 输入文本句子的特征提取是通过一个大小适中的过滤器来完成的,如公式所示:

$$\boldsymbol{c}_i = f\left(\mu * \boldsymbol{x}_{i:i+h-1} + \boldsymbol{b}\right) \tag{1}$$

其中,μ是卷积核;h表示卷积核的大小;$\boldsymbol{x}_{i:i+h-1}$是一个由$\{i: i+h-1\}$区间内词组成的嵌入;b表示偏差项;通过卷积层得到特征矩阵$\boldsymbol{v}=\left[\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \cdots, \boldsymbol{v}_{n-h+1}\right]$.

(3)最大池化层. 通过对卷积层得到的句子局部特征矩阵$\boldsymbol{s}$进行下采样,得到局部值的最优解$\boldsymbol{M}_i$. Max-Pooling技术的原理如式(2)所示:

$$\boldsymbol{M}_i = \max\left(\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_{n-h+1}\right) = \max\left(\boldsymbol{s}\right) \tag{2}$$

由于BiLSTM的输入必须是序列化的结构,而池化层将中断序列结构,因此有必要在池化层之后添加一个全连接层,将嵌入的$\boldsymbol{M}_i$连接到特征矩阵$\boldsymbol{U}$中.

3.3 BiLSTM

BiLSTM(Bi-directional Long Short-Term Memory)是一种循环神经网络模型,BiLSTM具有记忆前序状态的能力,比较适合处理序列问题,能更好的控制信息的流动,捕获更长的依赖关系且模型较为稳定.

BiLSTM由正向LSTM、反向LSTM、正向和反向LSTM的输出状态连接层组成,一个新的连续高阶矩阵$\boldsymbol{U}$被用作BiLSTM的输入. 如果t时刻正向LSTM输出的隐藏状态为$\overrightarrow{\boldsymbol{h}_t}$,反向LSTM输出的隐藏状态为$\overleftarrow{\boldsymbol{h}_t}$,则BiLSTM输出为隐藏状态$\boldsymbol{h}_t$,其具体计算过程如公式所示:

$$\overrightarrow{\boldsymbol{h}_t} = \overrightarrow{\mathrm{LSTM}}\left(\overrightarrow{\boldsymbol{h}_{t-1}}, \boldsymbol{U}_t\right) \tag{3}$$

$$\overleftarrow{\boldsymbol{h}_t} = \overleftarrow{\mathrm{LSTM}}\left(\overleftarrow{\boldsymbol{h}_{t-1}}, \boldsymbol{U}_t\right) \tag{4}$$

$$\boldsymbol{h}_t = \boldsymbol{\omega}_t\overrightarrow{\boldsymbol{h}_t} + \boldsymbol{v}_t\overleftarrow{\boldsymbol{h}_t} + \lambda b_t \tag{5}$$

其中,$\boldsymbol{\omega}_\tau$、$\boldsymbol{v}_t$为加权矩阵;$\boldsymbol{U}_t$是t时刻的LSTM输入;λb_t是偏移量. 本文通过正向LSTM层得到了隐层表示序列$\left(\overrightarrow{\boldsymbol{h}_1}, \overrightarrow{\boldsymbol{h}_2}, \cdots, \overrightarrow{\boldsymbol{h}_n}\right)$,通过反向LSTM得到隐层表示序列$\left(\overleftarrow{\boldsymbol{h}_1}, \overleftarrow{\boldsymbol{h}_2}, \cdots, \overleftarrow{\boldsymbol{h}_n}\right)$,最后通过正向和反向LSTM的输出状态连接层将两个隐层表示序列中单词对应的向量拼接,得到新的隐层序列表示$\left(\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n\right)$,其中$\boldsymbol{h}_i = \left(\overrightarrow{\boldsymbol{h}_i}, \overleftarrow{\boldsymbol{h}_i}\right)$. 利用双向LSTM并行信道的特性,本文提出的模型能够同时获得前向累积依赖信息和后向预测依赖信息,进而丰富了提取的特征信息. 设BiLSTM的隐层数为e,BiLSTM的输出结果公式如下:

$$\boldsymbol{\psi} = \left[\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \cdots, \boldsymbol{h}_n\right] \tag{6}$$

上式中,$\boldsymbol{\psi} \in \mathbf{R}^{n\times(2\times e)}$,$\boldsymbol{\psi}$中的每一行表示BiLSTM提取的字符级向量特征.

3.4 MUH-Attention机制

在我们的模型中,采用了结合ReCNN和BiLSTM网络的混合神经网络. 由于ReCNN输出数据的稀疏性,本文加入了多头注意力机制增强重要信息. 具体流程为:先对查询$\boldsymbol{Q}$(Query)、键$\boldsymbol{K}$(Key)和值$\boldsymbol{V}$(Value)进行线性变换,然后输入缩放点积注意力机制. 缩放点积注意力机制是一个使用点积进行相似度计算的注意力机制,并且参数是不共享的. 计算方式如式(7)所示.

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{V}}{\sqrt{d_k}})\boldsymbol{V} \tag{7}$$

其中，$\boldsymbol{Q}\in\mathbf{R}^{n\times d_k}$ 为查询矩阵，$\boldsymbol{K}\in\mathbf{R}^{n\times d_k}$ 为键矩阵，$\boldsymbol{V}\in\mathbf{R}^{n\times d_k}$ 为值矩阵，$\sqrt{d_k}$ 是调节因子，d_k 表示维度．在实验中重复 H 次上述操作，将 H 次标度点积的注意力结果拼接，并将线性变换得到的值作为 MUH-Attention 机制的结果[23]，得到特征信息．

$$\boldsymbol{H}_i=\text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}\right) \tag{8}$$

$$\text{MUH}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Concat}\left(\boldsymbol{H}_1,\cdots,\boldsymbol{H}_h\right)\omega^{o} \tag{9}$$

其中，$\boldsymbol{W}$ 是要训练的不同参数的权重参数矩阵，h 个头则对应 h 个维数，每个维数都是一个嵌入．$\boldsymbol{W}^{\boldsymbol{Q}}$ 是 $\boldsymbol{W}$ 矩阵的第 i 维．$\boldsymbol{Q}$，$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 的线性变换参数 $\boldsymbol{W}$ 均不相同．经过降维计算后，将输入嵌入后得到的向量直接赋给 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$，即 $\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{Z}$，最后将具有位置信息的嵌入字输入 BiLSTM：

$$\text{PE}_{(2i)}(p)=\sin(\text{pos}/10\ 000^{2i/d_{\text{model}}}) \tag{10}$$

$$\text{PE}_{(2i+1)}(p)=\cos(\text{pos}/10\ 000^{2i/d_{\text{model}}}) \tag{11}$$

在上述公式中，关于 p 的位置运算称为第 1 个元素 p 在 d_{model} 维中的位置嵌入运算．位置嵌入是绝对位置信息，但相对位置也很重要．上述位置嵌入公式的一个原因是 $\sin(\alpha+\beta)=\sin\alpha\cos\beta+\cos\alpha\sin\beta$ 和 $\cos(\alpha+\beta)=\cos\alpha\cos\beta-\sin\alpha\sin\beta$，这表明 $p+k$ 位置的嵌入可以表示为 p 位置嵌入的线性变换，为表达相对位置信息提供了可能．

3.5　空间域滑动窗口

由于生物医学文献的特殊性，语料中长文本数据较多，导致远程依赖问题十分突出．为此，本文提出了一种结合空间域滑动窗口和 Concept() 函数的方法来解决这一问题．空间域滑动窗口解决了由于将长文本分开所导致两个相邻的 batch 中出现相同的单词的问题．例如处理过的句子向量 $\boldsymbol{T}=\left[\boldsymbol{T}_1,\boldsymbol{T}_2,\boldsymbol{T}_3,\boldsymbol{T}_4,\boldsymbol{T}_5,\boldsymbol{T}_6,\boldsymbol{T}_7\right]$ 将被空间域滑动窗口划分为 $\boldsymbol{B}_1=\left[\boldsymbol{T}_1,\boldsymbol{T}_2,\boldsymbol{T}_3\right]$，$\boldsymbol{B}_2=\left[\boldsymbol{T}_3,\boldsymbol{T}_4,\boldsymbol{T}_5\right]$，$\boldsymbol{B}_3=\left[\boldsymbol{T}_5,\boldsymbol{T}_6,\boldsymbol{T}_7\right]$．然而，由于空间域滑动窗口将一句话切分为等长的片段，可能会导致上下文信息流失．为此，本文利用 Concept() 函数将上个 batch 的最后一个单词的词向量与下一 batch 的第一个单词的词向量进行运算，从而保证上下文的语义信息，本文将其定义为 Concept() 函数．本文通过空间域滑动窗口将 MLEE 特有的长句子分为等批次同时送进神经网络进行训练，有关空间域滑动窗口 Q_{special} 的公式如下：

$$\boldsymbol{R}_{\text{Related}}=\sum_{\substack{\lim\\ \text{Right},\text{Left}\to\infty}}\left(\boldsymbol{B}_{\text{Right}}\oplus\boldsymbol{B}_{\text{Left}}\right),\text{Right}\in\{5,10,15,\cdots,5n\},$$
$$\text{Left}\in\{6,11,16,\cdots,5n+1\},n\in\mathbf{N} \tag{12}$$

$$\text{Concept}(W)=\left[\text{count}\left(W\cdot\boldsymbol{R}_{\text{Related}}\left(\sum_{n=1}^{B=5}\left(\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{B}_3,\cdots,\boldsymbol{B}_n\right)\right)\right)\right],$$
$$n\in\boldsymbol{N} \tag{13}$$

$$Q_{\text{spacial}}=\lambda\nabla\cdot\text{Concept}(W) \tag{14}$$

其中，$\boldsymbol{B}_{\text{Right}}$ 表示上一批短句中最右边的单词，$\boldsymbol{B}_{\text{Left}}$ 表示下一批短句中最左边的单词，$\boldsymbol{R}_{\text{Related}}$ 表示针对以上两个单词创建的关系，W 表示通过 Concept() 函数连接相邻 batch 参数的运算结果，$\lambda\nabla$ 用于避免梯度的快速丢失．如算法 1 是 Concept() 函数的相关描述．

算法 1　Identify RowContext

输入: $\boldsymbol{B}_{\text{Right}}(T_i)=(T_1,T_2,\cdots,T_n)$ and $\boldsymbol{B}_{\text{Left}}(K_i)=(K_1,K_2,\cdots,K_n)$
输出: Concept(W)

```
1. Initialize: Quene q, j = 6, W
2.  for i = 1; j ≤ n;do
3.    if f( q.size( ) < 5) then
4.     q.push(B^(i)_Right)
5.    end
6.    else
7.     W = W + Σ_{k=1}^{k=q·since} [q(k_1,k_2,⋯,k_n) + B^(i)_Left], Concept(W)
8.     while q not empty do
9.       q.pop ()
10.     end
11.      j +=5;
12.   end
13. end
14. return Concept(W)
```

本文使用负对数似然函数作为损失函数，假设共有 N 个训练样本 $\left(\text{Tra}_i,\text{Tes}_i\right)$，则公式如下：

$$\text{Loss}\left(\chi\right)=-\sum_{j=1}^{N}\left[\log p\left(\text{Tra}_j,\text{Tes}_j\right),\log p\left(\chi\right)\right]+\lambda\|\theta\|_1 \tag{15}$$

在上述公式中，χ 是 $\log p\left(\chi\right)$ 在神经网络中的参数，是 L_1 正则化中的参数．上述函数使得参数稀疏，减少了计算开销，提高了模型计算速度．

4　实验

在本节内容中，本文将我们的模型与基线模型和其他几个模型进行了比较，描述了实验中使用的数据集，给出了超参数设置的细节，并给出了评价指标．此外，我们还详细分析了模型的有效性．

4.1　实验设置

4.1.1　数据集

本文实验采用的数据集为生物医学事件抽取通用语料 MLEE[6]．如表 1 所示，MLEE 语料库包含 295 篇生物医学文献摘要，总计包括 2 608 个句子和 6 677 个事件．训练集中包含 3 598 个触发词，测试集中有 1 809 个触发

词,触发词类别即为生物事件类型. MLEE语料库涵盖了从分子水平到生物体水平的各个层次的事件. 本语料库中的事件大致分为“Anatomical”、“Molecular”、“General”和“Planned”四大类,可进一步分为19个子类.

表1 数据集统计

数据名称	训练集	验证集	测试集	总计
文档数	206	30	59	295
句子数	1 825	260	523	2 608
事件	4 673	668	1 336	6 677

4.1.2 评价指标

MLEE语料库分为训练集、开发集和测试集三个部分. 我们将训练集和开发集合并送入模型进行训练,使用开发集调整参数,测试数据集进行测试. 利用P(精准度)、R(召回率)、F(F值)对所提出的方法进行了评价. 评价指标P、R、F由式(16)定义,其中TP、FP和FN分别表示真阳性、假阳性和假负性.

$$\begin{aligned} P&=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times 100\%,\\ R&=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\%,\\ F&=\frac{2\times P\times R}{P+R}\times 100\% \end{aligned} \tag{16}$$

4.1.3 实验参数设置

本文神经网络框架基于Pytorch[21]实现,经过近百次训练确定最终参数设置. 其中,BioBERT生成的词向量的维度是768维. 空间域滑动窗口从集合{2,3,4,5,6,7,8,9,10}选取了5. 神经网络参数shuffle设置为true,优化器采用随机梯度下降(SGD),在ReCNN和BiLSTM中的dropout值设置为0.5,学习率从集合$\{0.1,0.01,1\times10^{-3},1\times10^{-4},1\times10^{-5},1\times10^{-6}\}$中选取了$1\times10^{-5}$. 此外,我们通过网络爬虫技术在PubMed获取了5.7 GB的领域语料库使用数据增强技术将语料库扩充,从而减少过拟合情况的发生,参数详见表2.

表2 实验相关参数

参数	参数名	参数值
Wordvec dim	词向量维度	768
ReCNN+BiLSTM	残差卷积神经网络+双向长短期神经网络层数	8+10
Dropout rate	混合神经网络Dropout率	0.5
Learning rate	混合神经网络学习率	1×10^{-5}
Optimizer	混合神经网络优化器	SGD
Windows_size	空间域滑动窗口大小	5
Label schema	MLEE语料标注方法	BIO

4.2 实验结果与分析

为了避免传统方法在特征提取过程中对自然语言处理工具的过度依赖,进而降低人工成本,本文提出了一种由BiLSTM和ReCNN组成的混合神经网络. 然后将MUH-Attention机制融入混合神经网络,同时为词嵌入赋予相对位置信息,以提高生物医学触发识别的准确性. 针对语料库中长句子数量过多而导致的长距离依赖问题,我们采用了空间域滑动窗口,在保持上下文语义信息不被破坏的情况下,对长句子进行等份分割.

如表3所示,本文以BiLSTM作为实验的基线模型,其F值为72.57%. 然后,引入了ReCNN提取句子中每个单词的局部特征,从而提高每个句子的触发词识别准确率,F值提高了3.13%. 此外,为了突出句子中的关键信息,本文在混合神经网络(ReCNN-BiLSTM)的基础上,融合了MUH-Attention机制,F值提高了1.89%. 最后,通过空间域滑动窗口对文本进行处理,极大地解决了生物医学语料中长文本带来的远程依赖问题,F值提高了3.56%. 本文提出的基于MUH-Attention机制和空间域滑动窗口的混合神经网络触发词识别模型在生物事件抽取通用语料MLEE语料库上,F值达81.15%.

表3 实验模型的自身比较 单位:%

Methods	F
Baseline (BioBERT+BiLSTM)	72.57
BiLSTM+ReCNN	75.70
BiLSTM+ReCNN+MUH-Attention	77.59
BiLSTM+ReCNN+MUH-Attention+Spatial domain sliding window	**81.15**

5 实验比较与分析

在MLEE语料库上,已有一些值得借鉴的、先进的生物医学事件触发词识别方法. 在本节中,我们将本文提出模型与其他先进方法的实验结果进行了比较. 此外,为了研究本文所提出方法的潜力,我们也给出了详细的性能比较和误差分析.

5.1 总体性能比较

为了验证我们的模型,我们选择了以下先进的模型进行比较.

(1)Pyysalo等人的SVM方法[6]:他们使用支持向量机SVM方法,手工设计显著特征,在触发词识别中F值为75.84%.

(2)He等人的两阶段模型方法[7]:在之前工作中我们提出了基于SVM和PA算法的两阶段的传统方法,同时结合了特征选择方法,在MLEE语料库上的F值为79.75%.

(3)Zhou等人的领域知识识别方法[8]:他们使用Medline构建了一个大型文本语料库,并通过神经网络将其嵌入到单词特征中,F值为78.32%.

(4)Wang等人的CNN方法[15]:他们使用了一种基于依赖词嵌入的方法,从原始输入中学习重要特征,并对其进行分类,F值为78.27%.

(5)Shen等人的CHNN-ELM方法[14]:他们提出了

针对生物医学事件触发词识别的CHNN-ELM框架,使用CNN的输入端到输出端端监督方法自动学习语义进行特征选择,F值为80.57%.

(6)Nie等人的ANN-embeddings方法[9]:他们使用词嵌入来辅助神经网络预测模型进行事件触发识别,F值为77.23%.

(7)Diao等人的混合神经网络方法[13]:他们提出了一种由SVM和BiLSTM组成的混合神经网络,通过BiLSTM与细粒度分类的方法对生物医学事件触发词识别,通过SVM与SoftMax混合方法进行分类,F值为80.66%.

(8)Wei等人的多层BiLSTM神经网络方法[11]:他们引入语言模型来动态计算上下文化的单词表示,并提出了一种多层剩余双向长短时记忆(BiLSTM)结构,来解决词语表达的模糊性和浅层隐含层特征提取的不足的问题,F值为80.74%.

本文选择了Pyysalo等人[6]的实验结果作为基线方法. 如表4所示,相对于传统的触发词识别方法,大多数基于深度学习的触发词识别方法(文献[11,13,14]方法和本文方法)获得了更好的性能. 此外,本文的F值比基线方法高6.36%,比Nie等人[9]高4.97%,比Zhou等人[8]高3.88%,相较于Wei等人的多层BiLSTM神经网络方法[11],本文方法的F值提高0.41%,在一定程度上说明了混合神经网络的有效性. 与Diao等人的混合神经网络模型[13]相比,本文模型的F值高1.54%,进一步证明了空间域滑动窗口和MUH-Attention机制对于触发词识别性能提升的有效性. 上述先进方法各有其独特的优点. 然而,生物医学文献长文本中普遍存在的远距离依赖问题未得到较好处理,且长文本中不同单词的语义信息重要程度差异较大. 为了解决上述问题,我们构造了混合神经网络模型提高预测精度,并在模型中结合了空间域滑动窗口和MUH-Attention机制.

表4　本文模型与其他先进的模型的比较　　单位:%

Methods	P	R	F
Pyysalo et al.[6]	70.79	81.69	75.84
Nie et al.[9]	71.04	84.60	77.23
Wang et al.[15]	73.56	83.62	78.27
Zhou et al.[8]	75.56	81.29	78.32
He et al.[7]	82.79	76.56	79.75
Shen et al.[14]	80.06	81.25	80.57
Diao et al.[13]	80.03	81.54	80.66
Wei et al.[11]	79.89	81.61	80.74
Ours	83.48	80.89	81.15

5.2　详细性能比较

为了进一步讨论本文提出方法在生物医学事件触发词识别各类别的表现,本节对子类识别详细结果进行了分析. 表5将本文方法与其他文献提供的子类性能的模型在精确度、召回率和F值上的实验结果进行了对比. 在MLEE语料库中,虽然只有四种复杂事件类型,但它们占了近50%的比例. 因此,复杂事件的性能在一定程度上决定了事件抽取的整体性能. 如表5所示,在“Regulation”、“Positive_regulation”和“Negative_ regulation”复杂事件类型中,本文方法的F值相对较高,证明了本文提出的融合MUH-Attention机制的混合神经网络模型的有效性. 利用MUH-Attention机制的位置嵌入方法使得每个触发词都拥有各自准确的相对位置信息,有助于复杂事件的触发词识别.

表5　实验中每个子类别的F值对比　　单位:cm

Category	Method	P	R	F
Cell_proliferation	Ours	75.6	73.8	74.7
	Pyysalo[6]	63.8	69.8	66.7
	Zhou[8]	78.4	67.4	72.5
	Nie[9]	81.4	89.7	85.4
Transcription	Ours	60.0	60.0	60.0
	Pyysalo	25.0	14.3	18.2
	Zhou	0.0	0.0	0.0
	Nie	24.0	85.7	37.5
Catabolism	Ours	33.3	33.3	33.3
	Pyysalo	0.0	0.0	0.0
	Zhou	16.7	33.3	22.2
	Nie	12.5	25.0	16.7
Phosphorylation	Ours	33.3	33.3	33.3
	Pyysalo	50.0	100.0	66.7
	Zhou	75.0	100.0	85.7
	Nie	100.0	100.0	100.0
Dephosphorylation	Ours	0.0	0.0	0.0
	Pyysalo	0.0	0.0	0.0
	Zhou	100.0	100.0	100.0
	Nie	100.0	100.0	100.0
Localization	Ours	78.6	70.3	73.8
	Pyysalo	79.9	83.5	81.6
	Zhou	80.9	85.7	83.2
	Nie	65.5	84.2	73.7
Binding	Ours	81.6	71.4	76.2
	Pyysalo	84.0	76.4	80.0
	Zhou	81.1	78.2	79.6
	Nie	81.8	80.4	81.1
Regulation	Ours	87.1	74.9	80.6
	Pyysalo	46.5	60.4	52.5
	Zhou	56.5	53.1	54.7
	Nie	59.9	68.0	63.7
Positive_regulation	Ours	84.9	75.7	80.0
	Pyysalo	67.9	86.7	76.1
	Zhou	71.6	86.4	78.3
	Nie	67.1	91.0	77.3

续表

Category	Method	*P*	*R*	*F*
Negative_ regulation	Ours	89.1	80.8	84.7
	Pyysalo	74.4	77.0	75.7
	Zhou	77.1	78.8	78.0
	Nie	70.9	84.6	77.1
Planned_process	Ours	87.0.	82.0	84.0
	Pyysalo	53.9	75.0	62.7
	Zhou	56.5	75.6	64.7
	Nie	64.5	74.9	69.3
Development	Ours	88.0	84.0	86.0
	Pyysalo	68.1	83.5	75.0
	Zhou	69.3	81.4	74.9
	Nie	48.5	84.7	61.7
Blood_vessel_development	Ours	88.0	84.0	86.0
	Pyysalo	95.7	96.3	96.0
	Zhou	98.7	97.3	98.0
	Nie	96.6	93.1	94.8
Growth	Ours	89.0	80.0	84.0
	Pyysalo	69.1	83.9	75.8
	Zhou	77.1	83.9	80.3
	Nie	100.0	91.1	95.3
Death	Ours	89.0	92.0	91.0
	Pyysalo	56.9	94.3	71.0
	Zhou	72.1	88.6	79.5
	Nie	62.8	88.9	73.6
Breakdown	Ours	83.0	87.0	85.0
	Pyysalo	80.0	34.8	48.5
	Zhou	80.0	34.8	48.5
	Nie	84.2	69.6	76.2
Remodeling	Ours	80.0	80.0	80.0
	Pyysalo	85.7	60.0	70.6
	Zhou	85.7	60.0	70.6
	Nie	16.7	10.0	12.5
Synthesis	Ours	100.0	100.0	100.0
	Pyysalo	33.3	50.0	40.0
	Zhou	40.0	50.0	44.4
	Nie	75.0	75.0	75.0
Gene_expression	Ours	88.0	88.0	88.0
	Pyysalo	83.8	93.9	88.6
	Zhou	84.7	92.4	88.4
	Nie	85.2	91.7	88.3

5.3 误差分析

如图3所示,图中较暗的部分是*F*值较高的类型. 从图中可以看出,本文模型在大多数子类中都取得了较好的效果,特别是在复杂事件的“Regulation”(调节)、“Positive_regulation”(正调节)和“Negative_regulation”(负调节)类型,本文获得较高*F*值. 但对于一些简单事件类型,如“Precipitation”和“Catabolism”,文本方法*F*值相对较低. 根据预测结果,我们将主要误差原因分类如下:

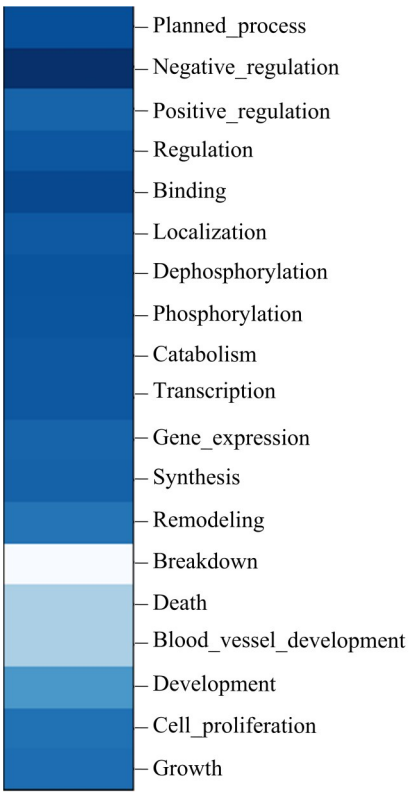

图3 触发词类别排列情况(颜色越深,*F*值越高)

(1)语料库中负例的数量较多,且正例类型的数量差异较大,数据不平衡问题明显.

(2)同一触发词在不同的训练集和测试集中的类别不同,导致了歧义.

(3)对于同一实体,事件中的含义可能不同. 例如,前文提到蛋白质的名称“PROTE-10”时,我们通常在下一句中用“it”代替“PROTE-10”,此类指代问题也对模型的识别造成了一定的误差性[9].

6 讨论

本文针对生物医学事件触发词识别中的长距离依赖问题和触发词的位置信息对触发词类型判断的影响,提出了融合空间域滑动窗口的混合神经网络嵌入MUH-Attention机制,在生物医学事件触发词识别中表现出良好的性能. 具体的优势分析如下:

(1)空间域滑动窗口

由于生物医学文献中存在大量的长句,句首的语义丰富而清晰,但随着句子长度的增加,句尾的语义丰富度降低,产生了长距离依赖问题. 因此,本文提出了一种空间域滑动窗口方法. 该方法在不破坏原始语义间上下文信息和位置信息的前提下,将生物医学文献中的长句划分为多个等长短句,形成窗口句子信息特征矩阵,有效地解决了触发词识别过程中由于长距离

依赖引起的重要信息丢失问题,进而提升了触发词识别性能.

(2)ReCNN-BiLSTM

本文提出了一种由ReCNN和BiLSTM组成的混合神经网络. 传统的触发词识别方法通过自然语言工具包来进行特征分类,这种方法不仅时间长还消耗大量的人工资源,模型泛化能力较弱;ReCNN与BiLSTM的结合有效地解决了传统方法在特征提取中过于依赖自然语言处理工具的问题,从而减少了人工成本. 同时综合利用了ReCNN和BiLSTM神经网络的优势. BiLSTM通过两个并行通道使触发词识别模型能够获得未来的累积依赖信息,从而丰富了提取的特征信息. ReCNN提高了单词级特征的识别准确率,加快了模型的收敛速度和损失值的下降速度. 由表2不难发现,混合神经网络的构建使触发词识别获得了明显的性能提升.

(3)MUH-Attention机制

触发词的位置信息以及上下文单词的语义信息直接影响着触发词类型的识别准确度. 因此本文引入了MUH-Attention机制,通过增加嵌入词的上下文语义信息,并利用其独特的位置嵌入方法为词嵌入提供更为准确的位置信息,从而提高触发词识别的准确性. 此外,通过注意力机制对每个单词都赋予权重,使得在距离触发词的相对距离较远的单词也能成功的被识别出来,进一步避免了长文本中重要信息丢失的问题.

7　结论及展望

本文提出了一种用于生物医学事件触发词识别的混合神经网络. 该模型利用ReCNN提取单词级特征,利用BiLSTM对上下文信息进行特征提取. 同时,采用MUH-Attention机制在词嵌入中添加更为精确的位置信息. 此外,本文还提出了一种新型的空间域滑动窗口方法,在不破坏上下文语义关系的前提下,将生物医学语料中特有的长文本分割成等长短文本批量地输入神经网络,在很大程度上解决了长距离依赖问题. 实验结果表明,本文提出的方法在生物医学事件抽取通用语料MLEE上达到了领先的性能.

在未来的工作中,我们将致力于利用优化模型的方法,以便解决生物医学事件抽取数据集中存在的数据稀疏问题. 同时,目前我们也在初步探索字符级神经网络方法,通过将单词拆解的方式,利用注意力机制,利用前后字母的组成信息,让生物医学事件触发词识别的准确率进一步提升.

参考文献

[1] GUPTA P, SCHUTZE H, ANDRASSY B. Table filling multi-task recurrent neural network for joint entity and relation extraction[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. Osaka: The COLING 2016 Organizing Committee, 2016: 2537-2547.

[2] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3: 1137-1155.

[3] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-01-16)[22-10-01]. http://arxiv.org/abs/1301.3781.

[4] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. New York: ACM, 2013: 3111-3119.

[5] MNIH A, HINTON G. A scalable hierarchical distributed language model[C]//Proceedings of the 21st International Conference on Neural Information Processing Systems. New York: ACM, 2008: 1081-1088.

[6] PYYSALO S, OHTA T, MIWA M, et al. Event extraction across multiple levels of biological organization[J]. Bioinformatics, 2012, 28(18): i575-i581.

[7] HE X Y, LI L S, LIU Y, et al. A two-stage biomedical event trigger detection method integrating feature selection and word embeddings[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018, 15(4): 1325-1332.

[8] ZHOU D Y, ZHONG D Y, HE Y L. Event trigger identification for biomedical events extraction using domain knowledge[J]. Bioinformatics, 2014, 30(11): 1587-1594.

[9] NIE Y F, RONG W G, ZHANG Y Y, et al. Embedding assisted prediction architecture for event trigger identification [J]. Journal of Bioinformatics and Computational Biology, 2015, 13(3): 1541001.

[10] RAHUL P V S S, SAHU S K, ANAND A. Biomedical event trigger identification using bidirectional recurrent neural network based models[EB/OL]. (2017-05-26)[2022-10-01]. http://arxiv.org/abs/1705.09516.

[11] WEI H, ZHOU A, ZHANG Y J, et al. Biomedical event trigger extraction based on multi-layer residual BiLSTM and contextualized word representations[J]. International Journal of Machine Learning and Cybernetics, 2022, 13 (3): 721-733.

[12] CHEN Y B, XU L H, LIU K, et al. Event extraction via dynamic multi-pooling convolutional neural networks [C]//Proceedings of the 53rd Annual Meeting of the Asso-

ciation for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2015: 167-176.

[13] DIAO Y F, LIN H F, YANG L, et al. FBSN: A hybrid fine-grained neural network for biomedical event trigger identification[J]. Neurocomputing, 2020, 381(C): 105-112.

[14] SHEN C, LIN H F, FAN X C, et al. Biomedical event trigger detection with convolutional highway neural network and extreme learning machine[J]. Applied Soft Computing, 2019, 84: 105661.

[15] WANG J, ZHANG J H, AN Y, et al. Biomedical event trigger detection by dependency-based word embedding [C]//Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Washington, DC: IEEE, 2015: 429-432.

[16] LI L S, HUANG M Z, LIU Y, et al. Contextual label sensitive gated network for biomedical event trigger extraction[J]. Journal of Biomedical Informatics, 2019, 95: 103221.

[17] 张亚军, 刘宗田, 周文. 基于深度信念网络的事件识别[J]. 电子学报, 2017, 45(6): 1415-1423.
ZHANG Y J, LIU Z T, ZHOU W. Event recognition based on deep belief network[J]. Acta Electronica Sinica, 2017, 45(6): 1415-1423. (in Chinese)

[18] SUN L, SONG Y. Research on Classifying performance of SVM with modified kernel function in HCCR[C]// ICNN&B'05, International Conference on Neural Networks and Brain. Beijing: IEEE, 2005, (3): 1720-1723.

[19] LI Y, YANG T. Word Embedding for understanding natural language: A survey[C]//Guide to Big Data Applications. Cham: Springer, 2018: 83-104.

[20] 潘博, 于重重, 张青川, 等. 基于词性与词序的相关因子训练的 word2vec 改进模型[J]. 电子学报, 2018, 46(8): 1976-1982.
PAN B, YU C C, ZHANG Q C, et al. The improved model for word2vec based on part of speech and word order [J]. Acta Electronica Sinica, 2018, 46(8): 1976-1982. (in Chinese)

[21] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.

[22] LEE J, YOON W, KIM S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12) [2022-10-01]. http://arxiv.org/abs/1706.03762.

[24] KETKAR N. Introduction to PyTorch[M]//Deep Learning with Python. Berkeley, CA: Apress, 2017: 195-208.

作者简介

任永功 男, 1972 年出生. 博士, 教授, 博士生导师, 现为辽宁师范大学计算机与信息技术学院教授. 主要研究方向为人工智能技术、数据库及数据挖掘技术等. 在《计算机学报》等国际、国内计算机类核心刊物上发表论文 60 余篇, 被 SCI、EI、ISTP 收录 30 余篇.
E-mail: ygren@lnnu.edu.cn

林禹竹 女, 1996 年出生, 辽宁师范大学计算机与人工智能学院在读博士, 主要研究方向为数据挖掘与智能计算、智慧教育与大数据技术.
E-mail: 46314368@qq.com

唐玉洁 女, 1999 年出生于辽宁省大连市. 现为辽宁师范大学硕士研究生. 主要研究方向为自然语言处理、文本挖掘和生物信息学.
E-mail: 609500343@qq.com

于 博 男, 1996 年出生. 2022 年毕业于辽宁师范大学, 获得理学硕士学位. 主要研究方向为数据挖掘、自然语言处理、文本挖掘和生物信息学.
E-mail: yubochina@aliyun.com

何馨宇 女, 1983 年出生. 博士, 副教授, 硕士生导师, 现为辽宁师范大学计算机与信息技术学院计算机科学与技术(师范)专业专任教师. 主要研究方向为文本挖掘和生物信息学. 目前已发表 30 余篇高水平学术论文, 其中第一作者 SCI、EI 检索十余篇, 包括国际顶级期刊 IEEE Trans 系列 TCBB(CCF 推荐 B 类 SCI 期刊)和领域顶级会议 BIBM(CCF 推荐 B 类会议)等.
E-mail: hexinyu@lnnu.edu.cn