基于实体信息和图神经网络的药物相互作用关系抽取

杨 霞[1]，韩春燕[2]，琚生根[1]
(1. 四川大学计算机学院，成都 610065；2. 四川民族学院理工学院，康定 626001)

摘 要：药物相互作用是指药物与药物之间相互促进或抑制. 针对现有的药物关系抽取方法利用外部背景知识和自然语言处理工具导致错误传播和积累的问题，以及现有大多数研究在数据预处理阶段对药物实体进行盲化，忽略了有助于识别关系类别的目标药物实体信息的问题. 论文提出了基于预训练生物医学语言模型和词汇图神经网络的药物相互作用关系抽取模型，该模型通过预训练语言模型获得句子的原始特征表示，在基于数据集构建的词汇图上进行卷积操作获得与句子相关的全局特征信息表示，最后与药物目标实体对特征进行拼接从而构建药物相互作用关系提取任务的特征表示，在获得丰富的全局特征信息的同时避免了使用自然语言处理工具和外部背景知识，提升模型的准确率. 论文的模型在 DDIExtraction 2013 数据集上的 F_1 值达到了 83.25%，优于目前最新方法 2.35%.

关键词：药物-药物相互作用关系抽取；预训练生物医学语言模型；目标药物实体对；图神经网络
中图分类号：TP391 **文献标识码**：A DOI：10.19907/j.0490-6756.2022.022002

Drug-Drug relationship extraction based on entity information and graph neural networks

YANG Xia[1], *HAN Chun-Yan*[2], *JU Sheng-Gen*[1]
(1. College of Computer Science, Sichuan University, Chengdu 610065, China;
2. College of Science and Technology, Sichuan University for Nationalities, Kangding 626001, China)

Abstract: Drug-Drug interaction refers to the mutual promotion or inhibition between drugs. For the existing drug relationship extraction methods, the use of external background knowledge and natural language processing tools leads to the problem of error propagation and accumulation, and most existing studies blind drug entities at the data pre-processing stage, ignoring the target drug entity information that is helpful to identify the relationship category. In this paper, a drug interaction extraction model based on pre-trained biomedical language model and word map neural network is proposed. In this model, the original feature representation of sentences is obtained by pre-trained language model, and the global feature information representation of sentences is obtained by convolution operation on the word map constructed based on data set. Finally, the feature representation of drug interaction relationship extraction task was constructed by stitching the feature with drug target entities, which can not only obtain rich global feature information but also avoid using natural language processing tools and external background knowledge, and improve the accuracy of the model. The F_1 value of the model on the DDIE-

收稿日期：2021-09-29
基金项目：国家自然科学基金(61972270)；四川省重点研发项目(2019YFG0521)
作者简介：杨霞(1992—)，女，四川宜宾人，硕士研究生，研究方向为知识图谱. E-mail：yxia_vip@163.com
通讯作者：琚生根. E-mail：jsg@scu.edu.cn

xtraction 2013 dataset achieved 83.25%, which outperforms the current latest methods by 2.35%.
Keywords: Drug-Drug interaction relationship; Pre-trained biomedical pretrained language model; Drug entity embedding; Graph neural network

1 引　言

药物-药物相互作用(Drug-Drug Interaction, DDI)关系抽取是生物医学关系抽取中最典型的任务之一,旨在从生物医学文献中提取两种或多种药物实体之间的相互作用关系. 在临床应用中,当多种药物同时服用时可能会发生药物相互作用,这种作用可能在增加或减少药物效果的同时,让服用者产生不良反应,医务人员往往花费大量时间审查DDI 的相关知识信息. 然而,随着生物医学文献数量的增加,手动收集 DDI 信息既费时又昂贵. 因此,如何有效地从这些医学文献中自动提取结构化信息已成为研究人员亟待解决的问题.

近年来,随着 DDIExtraction 2011[1] 和 DDIExtraction 2013[2] 药物抽取任务的发布,各种 DDI 提取方法被提出来,大致可以分为以下 3 类:基于模式匹配的方法、基于特征的机器学习方法和基于深度学习的方法. 基于模式匹配的方法是这一领域的传统方法, 利用特定类型的模式和匹配规则来识别生物医学实体之间的语义关系. 通常不需要标记数据,但需要生物医学专家来制定和设计模式形式或手动编码规则. 由于预先定义的模式或规则通常不能适应自由文本中的语法变化导致召回率较低. 因此,产生了机器学习的方法,基于机器学习的关系抽取采用特征表示或内核设计方法,通常会利用句子中多种多样的特征,并将其馈入支持向量机[3] 等分类器中. 与过去基于模式匹配的方法相比,基于特征的机器学习方法取得了较大的成功,并且具有更好的可移植性. 但是它仍旧需要人工定义特征,比如词性、句法、语法等. 由于获取这些特征需要利用外部自然语言处理(Natrual Language Processing,NLP)工具,而这些工具并非为特定领域量身定做,因此会存在错误传播从而影响性能. 得益于深度学习的兴起,基于神经网络的自动特征表示模型在 DDI 提取任务中取得了较大的成功,这些模型能够在没有大量手工特征工程的情况下,自动从训练数据中学习相关表示和特征,而无需专家仔细设计模式、特征和内核功能. 例如,卷积神经网络(Convolutional Neural Network ,CNN)和循环神经网络(Recurrent Neural Network,RNN). 然而,这些神经网络模型仅从给定的句子中提取语义特征,性能往往不能优于基于特征和内核方法的模型. 因此研究人员利用外部资源和背景知识来丰富语义特征,提升任务性能. 这些方法都极大地促进了 DDI 的抽取,但仍然存在几个缺陷. 首先,使用背景知识的模型可能过于局限于某些语料库,因为背景知识往往以不同的形式出现,有时候甚至找不到合适的知识. 其次,为了预测句子中药物实体对之间的相互作用关系,大多数方法除了利用句子中词汇信息外,还需要大量额外的特征,比如词性特征、句子的依赖特征以及语法树特征,而这些特征的提取依赖于 NLP 工具,因此可能会因为遭受错误传播和积累而导致实验性能下降.

受到 VGCN-BERT[4] 和预训练的生物医学语言模型(Biomedical Bidirectional Encoder Representations from Transformers, BioBERT)[5] 的启发,针对上述存在的问题,本文提出了基于预训练生物医学语言模型的词汇图卷积神经网络关系抽取模型(Relational BioBERT Vocabulary Graph Convolutional Network, RBio-VGCN),该模型通过 BioBERT 自动获得句子和实体嵌入特征,基于数据集中词语共现频率构建的词汇图,将句子嵌入与词汇图进行图卷积(Graph Convolutional NetWork, GCN)[6] 操作获得与句子相关的全局语义特征,通过 BioBERT 模型各个层中的自注意力机制将句子嵌入信息与全局语义相关信息充分交互,捕获与输入句子相关的信息并且忽略掉不相关的信息,得到与关系抽取任务相关的特征表示,最后与药物目标实体对特征进行拼接用于 DDI 关系抽取. 在获得较好的性能同时避免了使用外部资源和第三方 NLP 工具,使得该模型具有较好的泛化能力.

本文的主要贡献可归纳如下:(1) 首次在 DDI 数据集上构建词汇图,并将 BioBERT 获得的句子上下文信息使用图卷积神经网络获得与句子相关的全局特征,而不需要使用外部自然语言处理工具,避免错误传播与积累,最后使用多层自注意力机制,最大化获取与 DDI 任务相关的特征表示;(2) 通过在句子中嵌入目标药物实体对信息,为 DDI 关系抽取提供丰富的特征信息,而先前大多数

工作都将其进行盲化处理；(3) 模型在数据集 DDIExtraction 2013 上获得了最优结果，验证了该模型的有效性.

2　相关工作

目前在药物相互作用关系抽取领域应用的方法主要分为：基于模式匹配、基于核函数和基于深度学习的方法. 其中，基于深度学习的方法由于可以自动地捕获输入句子的特征，实现药物相互关系自动抽取，已成为现在的研究热点. 基于模式匹配和基于核函数的方法需要使用大量事先定义的特征，如词性、语义、药物名等特征来完成对药物关系的抽取. Tomas 等[7]使用基于多数投票机制的核函数方法. Zheng 等[8]使用基于等价类和综合上下文信息的图内核. 一般来说，这些基于特征和内核的方法都严重依赖于设计精良的特征或核函数.

随着深度学习的发展，Liu 等[9]提出了基于句子依赖解析的卷积神经网络模型，由于 CNN 模型忽略了句法信息以及句子中单词之间的长距离依赖关系，该模型利用依存解析树来捕获这些信息，其中边表示两个单词之间的句法依赖. Zhao 等[10]提出了一种语法卷积神经网络，它结合了基于语法嵌入的特征和传统特征，以获得更好的表示. 为了识别句法信息，他们使用解析器生成谓词-自变量结构中的最短路径序列，而非传统的线性单词序列. 刘宁宁等[11]提出了基于胶囊网络的药物关系抽取方法，该方法首先根据原语句解析出两个药物之间的最短依存路径，利用双向长短期记忆网络分别获取原语句和最短依存路径的低层语义表示，结合胶囊网络进行药物相互抽取. 得益于图神经网络的发展，GCN 已经被成功应用于在任意图结构上，包括知识图谱、社交网络、依赖图等. Park 等[12]提出了图卷积网络注意力模型，采用基于注意力的修剪策略获得输入句子的上下文信息和句子的结构信息. 但是以上模型使用额外工具解析句子依赖构建图结构，使得该模型可能会遭受错误传播和积累，并且都忽略了药物目标实体对特征信息进而将其盲化处理.

基于注意力机制的模型被广泛应用于自然语言处理，其中预训练语言模型 BERT[13]由于其多层双向 Transformer[14]结构，利用多层多头注意力机制将句子的上下文信息从前向和后向集成到单词向量中. BioBERT 是第一个在生物医学领域语料库上经过预训练的语言表示模型，该模型使用 BERT 的权重作初始化参数，然后在生物医学领域的语料库 PubMed 摘要和 PubMed Central 全文本上进行训练. Nguyen 等[15]基于 Relation BERT[16]模型，使用 BioBERT 获得句子上下文信息，在药物相互作用数据集上取得了良好的性能. 但是该模型中将目标药物实体对盲化，并未使用目标实体对特征信息用于关系抽取. Zhu 等[17]使用 BioBERT 获得句子的嵌入，并且利用药物实体特征信息，但同时也从知识库中引入大量药物解释知识，用以解释说明数据集中药物特征信息. 由于需要引入特定的背景知识，降低了模型的泛化能力.

以上方法都严重依赖于语言特征和领域背景知识，这可能会给模型带来额外的错误以及影响模型的泛化能力. 并且目标实体在句子中的位置信息与目标实体的上下文语义信息，对于关系抽取具有促进作用，而先前大多数工作都在数据预处理阶段将药物实体盲化. 因此，本文提出的 RBio-VGCN 模型通过在 DDI 数据集上构建词汇图神经网络，使得句子获得额外信息特征的同时避免引入大量背景知识，同时利用数据集中的词汇构建图，避免使用第三方 NLP 工具解析句子依赖.

3　本文方法

3.1　任务描述

DDI 关系抽取是根据生物医学文献中的句子对两个药物实体之间的相互作用类型进行分类. 本文使用药物-药物相互作用公共数据集 DDIExtraction 2013 进行实验，药物-药物相互作用关系抽取实例如图 1 所示.

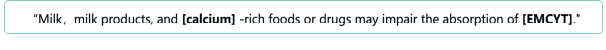

图 1　药物-药物相互作用关系抽取示例
Fig. 1　Examples of drug-drug interaction extraction

对于句子中给定的药物实体标记：e_1 =“calcium” 和 e_2 =“EMCYT”本文的目标是自动识别出句子中药物实体e_1和药物e_2所表达的关系 Mechanism.

3.2　模型简介

对于给定标记了目标实体e_1和e_2两个实体的句子 $s=[w_i]_{i=1}^{t}$，其中，t 表示句子的长度，w_i表示句子中的第 i 个词. 为了获取药物目标实体在句子中的位置信息，本文采用 Relation BERT 的方法，在目标药物实体e_1和e_2前后分别添加特殊标记“$”和“#”. 将标记后的句子输入模型，分别获得

句子中词的嵌入与实体e_1和e_2的嵌入. 为了获得与句子相关的全局特征信息,本文基于数据集中的词汇创建了词汇图,并将句子的原始嵌入与构建的词汇图进行卷积操作,然后使用 BioBERT 的多层注意力将句子特征与句子相关全局特征进行充分交互,尽可能捕获与 DDI 任务相关的句子特征信息,摒弃不相关的信息. 最后,将句子嵌入、句子相关全局嵌入和药物目标实体嵌入拼接后使用 Softmax 函数归一化后输出各个类别的概率,框架如图 2 所示.

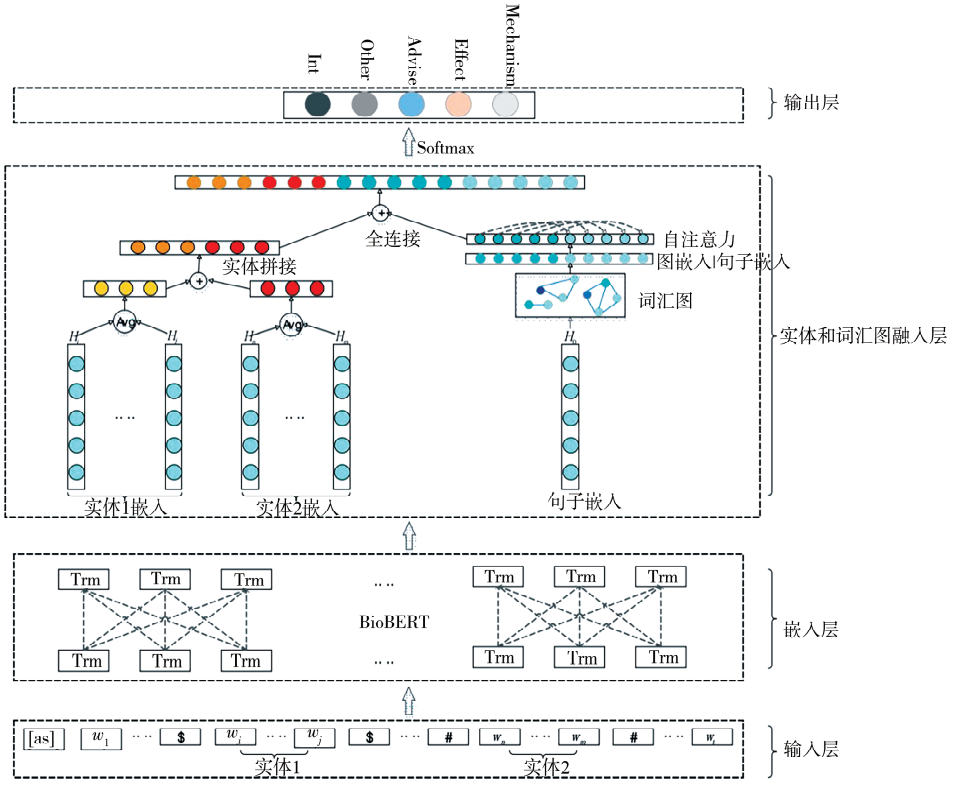

图 2　RBio-VGCN 模型结构

Fig. 2　Structure of Rbio-VGCN model

3.3　输入层

本文根据原数据中的药物实体,生成相互作用的药物实体对. 针对目标药物实体对不盲化,对句子中的非目标实体对的药物实体使用“GRUG0”进行盲化. 由于本文使用的数据集中已经标记好了药物实体,因此不再需要进行命名实体识别. 假设原语句 s 为:“Dexamethasone at 10(−10)M or retinyl acetate at about 3X 10(−9)M inhibits proliferation stimulated by EGF.”,其中“Dexamethasone”,“retinyl”以及“EGF”表示药物实体,该句中共有三个药物实体,经过药物实体两两组合之后,可以得到三组药物对句子,在对每个句子中的目标实体对进行特殊符号标记后,会产生三个输入语句. 如表 1 所示.

3.4　嵌入层

嵌入层是为了将句子中的词语映射到同一个语义空间中,并且根据句子的上下文语境将句子中的词语编码成向量,本文选择 BioBERT 对句子中的词语进行编码. 假设句子 $s=[w_i]_{i=1}^{t}$,其中 t 表示句子的长度,w_i 表示句子中的第 i 个词. 使用 BioBERT 对句子 s 进行上下文编码,得到嵌入表示s',嵌入公式(1)如下:

$$s'=[w_i']_{i=1}^{t}=\text{BioBERT}([w_i]_{i=1}^{t}) \tag{1}$$

其中,w_i'是词w_i的上下文编码向量,维度为$R^{d\times d}$,其中 d 是预训练语言模型隐藏层的维度.

表 1　输入语句处理

Tab. 1　Processing of input sentence

句子	药物对	类别	语句
s1	(Dexamethasone, retinyl)	false	$ Dexamethasone $ at 10(−10) M or # retinyl # acetate at about 3 X 10(−9) M inhibits proliferation stimulated by GRUG0.
s2	(Dexamethasone, EGF)	false	$ Dexamethasone $ at 10(−10) M or GRUG0 acetate at about 3X 10(−9) M inhibits proliferation stimulated by # EGF #.
s3	(retinyl, EGF)	effect	GRUG0 at 10(−10) M or $ retinyl $ acetate at about 3X 10(−9) M inhibits proliferation stimulated by # EGF #.

对于句子中的目标实体嵌入，本文将组成该目标实体的词的嵌入表示进行平均化，然后将平均后的结果作为该目标实体的嵌入表示，目标实体e_1和e_2嵌入公式分别如式(2)和式(3)所示.

$$e_1 = \frac{1}{j-i+1}\sum_{k=i}^{j} w'_k \tag{2}$$

$$e_2 = \frac{1}{m-n+1}\sum_{k=n}^{m} w'_k \tag{3}$$

其中，i、j 分别表示实体e_1中第 i 和第 j 个词语；m、n 分别表示实体e_2中第 m 和第 n 个词语.

3.5　实体和词汇图融入层

当嵌入层获得输入句子中词语的嵌入之后，在词汇图上进行卷积操作生成与句子相关的全局图嵌入，在此过程中，只有与输入句子相关的特征信息才会被抽取并且嵌入，再将句子嵌入和全局相关特征嵌入拼接(1)，对拼接后的特征表示信息使用 BioBERT 中的多层自注意力机制，让句子嵌入和句子全局图嵌入特征进行充分交互(2)，使得原始句子融入词汇中全局特征信息表示(3)，过程如图 3 所示.

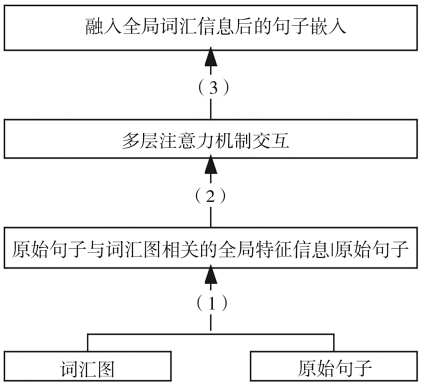

图 3　句子嵌入与词汇图交互过程

Fig. 3　Process of sentence embedding interacts with vocabulary graph

3.5.1　*构建词汇图*　本文使用标准点互信息(Normalized Point-wise Mutual Information，NPMI)构建词汇图，因为这个指标可以很好的衡量两个词语之间的相关性，如式(4).

$$\mathrm{NPMI}(i,j) = -\frac{1}{\log p(i,j)}\log\frac{p(i,j)}{p(i)p(j)} \tag{4}$$

其中，i 和 j 是词语；$p(i)$和 $p(j)$表示的是两个单词出现的频率；$p(i,j)$表示词语 i 和词语 j 在同一条句子中出现的概率. NPMI 的值的范围是[−1，1]，正数表示单词之间的语义相关性很高，而负数则表示很少或根本不相关. 在本文提出的方法中，如果两个单词之间的 NPMI 大于阈值，则在这两个单词之间建立一条边. 本文实验表明，当阈值在 0 到 0.2 之间时性能达到最优.

3.5.2　*词汇图卷积神经网络*　GCN 由 Kipf 等[6]提出，是一个直接在图上进行卷积操作的神经网络，通过邻居节点的属性推导当前节点的嵌入特征，从而在一定程度上集成该数据域的全局上下文信息. 给定一个单层的 GCN，卷积的过程如式(5)所示.

$$H = \widetilde{A}XW \tag{5}$$

其中，$X \in R^{n\times m}$表示 n 个节点 m 维特征的输入矩阵；$W \in R^{m\times h}$是权重矩阵；$\widetilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$是归一化对称对角矩阵，$A \in R^{|v|\times|v|}$，表示一条句子中以每个词语为顶点构成的邻接矩阵，对 A 进行归一化操作是为了避免深度神经网络模型梯度消失或爆炸的现象.

本文的目标是使用与任务相关的词语进行 DDI 关系抽取，而不是使用语料库中的整个句子，因此，本文提出的图形是基于词语构建的. 假设给定由 x 个单词组成的输入句子，可以用式(6)来表示单层图卷积的过程.

$$h = (\widetilde{A}x^T)^T W = x\widetilde{A}W \tag{6}$$

其中，$\widetilde{A}^T = \widetilde{A}$表示词汇图，$x\widetilde{A}$表示抽取出输入句子 x 与词汇图中词汇相关的全局特征. $W \in^{|V|\times h}$表示当前输入句子 x 的权重. 因此对于一个包含 m 条句子的训练批次，可以将式(6)转化为

$$H = X\widetilde{A}W \tag{7}$$

那么相应的一个带有激活函数的多层词汇图卷积网络可以表示为

$$\mathrm{VGCN} = \mathrm{ReLU}(X_{mv}\widetilde{A}_{vv}W_{vh})W_{hc} \tag{8}$$

其中，m 表示一个批次的句子条数；v 是训练集中词汇的数量；h 是隐藏层大小；c 是句子的嵌入维度；通过$X_{mv}\widetilde{A}_{vv}$捕获输入句子与词汇图神经网络相关的特征，最后通过图卷积操作获得与输入句子相关的全局词汇特征.

3.5.3　*多层自注意力机制*　在获得了与输入句子相关的全局词汇特征后，通过自注意力机制可以将原始句子的特征与全局词汇特征进行充分交互，在保留当前句子的上下文信息时融入与关系抽取相关的背景知识信息.

通过给定一个和任务相关的查询向量 Q，计算与 K 的注意力分数并附加在 V 上，从而计算注意力分数，使用注意力分数，每个词语可以获得一个矢量表示来编码上下文信息. 注意力分数计算公式如下：

$$\text{Attention}(Q,K,V)=\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

其中,$\sqrt{d_k}$是缩放因子,用于调整注意力分数的大小.

本文将输入句子的原始嵌入和词语图神经网络相关的嵌入和一同输入 BioBERT 中,不仅获得了词语在句子中的序列信息,还获得了从 VGCN 中捕获到的背景知识. 通过自注意力机制,将句子的局部信息和词汇图神经网络的全局信息进行充分交互,得到与任务相关的最终特征表示.

$$F_{\text{emb}}=ReLU(X_{\text{mev}}\widetilde{A}_{\text{vv}}W_{\text{vh}})W_{\text{hg}} \quad (10)$$

其中,W_{hg}中 g 是超参数,为词语图神经网络的输出维度;m 是一个训练批次的大小;e 是词语的嵌入维度;v 是训练集中词汇的数量.

在嵌入层分别获得标记后句子的表示输出 F_{emb}、目标实体e_1和目标实体e_2的表示输出后,本文对这三个输出表示进行拼接,得到最终的句子表示 H_f. 其中,权重矩阵w_1的维度为$R^{n\times 3d}$,b_1为偏置向量. 如式(11)所示.

$$H_f=W_1[\text{concat}(F_{\text{emb}},e_1,e_2)]+b_1 \quad (11)$$

3.6　输出层

输出层的作用将句子与目标实体全连接后的表示信息进行归一化,输出概率最大的标签. 输出层利用 Softmax 函数实现归一化,使得所有关系类别的总概率和为 1,如式(12)和式(13)所示.

$$\hat{y_i}=\text{Softmax}(H_f) \quad (12)$$

$$C^*=\text{argmax}(\hat{y_i}) \quad (13)$$

其中,H_f为输入句子和目标药物实体对在经过模型训练后的最终特征表示;$\hat{y_i}$为各个药物类别的概率;C^*为概率最大的关系类别.

本文采用交叉熵损失函数进行训练,交叉熵计算得分,该得分可以得出所有类别的实际概率分布与预测概率分布之间的差异.

$$Loss=-\sum_{i=1}^{n}\log p(y_i\mid s_i) \quad (14)$$

其中,n 表示训练数据集 $D=(\{s_1,y_1\},\cdots,\{s_n,y_n\})$中数据大小;$y_i$表示第 i 条句子s_i的真实类别标签;$\log(y_i|s_i)$表示第 i 条句子s_i被模型预测为真实标签y_i的概率.

4　实　验

4.1　数据集及评估函数

本文使用的药物-药物相互作用关系抽取数据集 DDIExtraction2013 如表 2 所示. 由 730 篇 DrugBank 中的医学文本和 MEDLINE 中的 175 篇摘要组成. 该数据集分成两部分:训练集由 572 篇 DrugBank 中的医学文本和 MEDLINE 中的 142 篇摘要组成;测试集由 158 篇 DrugBank 中的医学文本和 MEDLINE 中的 33 篇摘要组成. 该数据集中所有的药物实体都进行了标注,共有以下 5 种药物-药物相互作用关系类型. 1) Mechanism:描述两种药物实体的药代动力学机制;2) Effect:明确地指出了两种药物相互作用的结果;3) Advise:描述了两种药物同时使用时的建议;4) Int:说明两种药物存在一定的关系,但未定义具体的关系类型;5) Negative:说明两个药物之间不存在相互作用.

表 2　DDIExtraction 2013 数据集信息统计

Tab. 2　The statistics information of DDIExtraction 2013 dataset

类型	训练集	测试集	总数
Negative	23 772	4737	28 509
Mechanism	1319	302	1621
Effect	1687	360	2047
Advise	826	221	1047
Int	188	96	284

现有的 DDI 提取模型采用召回率 R(Recall)、精确率 P(Precision)、F_1值(F_1-score)三个指标进行评估.

4.2　参数设置

本文实验条件为 1 个 RTX3090-24G,使用 PyTorch 框架(https://github.com/pytorch/pytorch),预训练语言模型采用基于医学数据集上进行训练的 BioBERT,该模型包含 12 层的 Transformer. 本文模型使用的参数取值如表 3 所示.

表 3　参数取值

Tab. 3　Parameter value

参数	取值
Epoch	20
Batchsize	16
学习率	2e−5
最大句子长度	200
句子嵌入层维度	768
图卷积网络嵌入层维度	16
Dropout	0.1

4.3　实验结果

本文模型在 DDIExtraction2013 数据集上的训练过程如图 4 所示,图 4 是模型的 F_1值曲线图.

从图中可以看出，模型训练的前段部分 F_1 值提升较快，后续不断的波动寻找局部最优值，最后逐渐趋近平稳.

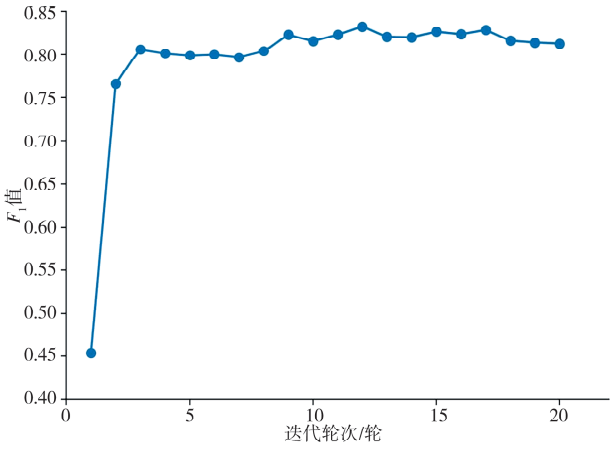

图 4　F_1 值曲线图
Fig. 4　F_1 value graph

本文设计了消融实验，以便于更好的分析不同模块对 DDI 抽取的影响力. 如表 4 所示，分别验证了词汇图卷积神经网络（VGCN）和嵌入目标药物实体特征（RBio-BERT ＋ Entity）对实验结果的影响.

由表 4 消融实验可知，当仅使用预训练的 BioBERT 模型而不加入任何实体信息和词汇图神经网络时，模型并不能很好的识别 DDI 关系. 在加入目标实体信息后，模型的效果提高了 1.55%，说明目标实体信息有利于关系抽取实验性能. 而在加入原始 BioBERT 模型上加入词汇图神经网络捕获句子全局特征之后，实验性能提升了 1.15%，说明与句子相关的全局信息可以提升关系抽取准确率.

表 4　消融实验
Tab. 4　Ablations experiment

模型	P/%	R /%	F_1/%
RBio-BERT ＋VGCN＋ Entity	82.49	84.02	83.25
RBio-BERT ＋ Entity	81.95	81.86	81.90
RBio-BERT ＋VGCN	81.12	81.86	81.49
RBio-BERT	82.66	78.17	80.35

该消融实验的结果说明本文提出的模型可以充分结合预训练生物医学语言模型、词汇图神经网络、目标药物实体信息三者的优势，从而更好地提升整个模型的抽取效果.

为了更好地验证本文模型的有效性，本小节将 RBio-VGCN 模型的性能与该数据集的其他模型[11,15,17-22]进行了比较. 表 5 展示了在数据集 DDIExtraction 2013 上不同模型的实验结果. 从表 5 中可以看到本文模型在测试集上的结果分别为：F_1 为 83.25%，P 为 82.49%，R 为 84.02%，并且每种 DDI 类型的 F_1 值也是优于先前的工作. 在比较了现有最新模型之后，本文提出模型的 F_1 值比现有最好模型[17]高出 2.35%.

表 5　基线模型实验结果比较
Tab. 5　Comparison of baseline model experimental results

模型	每种 DDI 类型的 F_1 值				模型的整体性能		
	Advise	Effect	Mechanism	Int	P/%	R /%	F_1/%
MCCNN[18]	77.97	68.23	72.2	51.3	75.99	65.25	70.21
CNN-GCNs[19]	81.62	71.03	73.83	45.83	73.31	71.81	72.55
RHCNN[20]	80.5	73.4	78.2	58.9	77.30	73.75	75.48
PM-BLSTM[21]	81.60	71.28	74.42	48.57	75.80	70.38	72.99
BiLSTM-CapsNet[11]	—	—	—	—	78.78	70.5	74.16
ATT-BLSTM[22]	85.1	76.6	77.5	57.7	78.40	76.20	77.30
R-BioBERT[15]	87.32	97.42	77.80	57.31	—	—	80.89
BioBERT+BiGRU[17]	86.0	80.1	77.5	56.6	81.00	80.90	80.90
Ours	88.74	81.41	87.52	59.21	82.49	84.02	83.25

5　实验分析

5.1　药物目标实体和词汇图神经网络对实验性能的影响

模型在加入目标药物实体和词汇图神经网络之后的精确率、召回率和 F_1 值如图 5 所示. 从图中可以得出,在加入该信息之后,实验性能提升了 2.9%,这充分说明了药物实体对于信息 DDI 分类是有促进作用的. 而对于本文构建的词汇图,在通过图卷积神经网络更新节点特征获得与句子相关的特征信息后,充分利用自注意力机制获得与 DDI 分类相关的特征,使得分类准确率得到了提升,也充分论证了全局信息对于实验性能提升是有促进作用的.

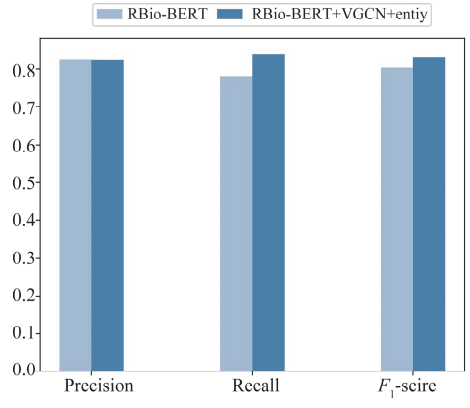

图 5　药物实体和词汇图对实验性能的影响

Fig. 5　Effects of drug entities and vocabulary graph on test performance

5.2　模型错误分类分析

图 6 是本文模型的混淆矩阵,图中颜色越深表示所占的比例越大. 为了突出模型对药物关系类别的错误分类,本文将每一种 DDI 类别的数量进行归一化处理. 从图 6 可以看出,该模型分类错误主要有两种:1) 类别为 Int 的这一类关系经常被错误分类为 Effect 类;2) 四种正例关系类别(Effect,Mechanism,Advise,Int)经常被错误地分类到负例这一类别中.

第一种类型的错误分类和先前的一些工作非常类似[23]. 我们认为原因在于 Int 类型的数量太少,训练集中仅有 96 条实例,并且本文观察到数据集中类型为 Int 和 Effect 的实例具有相似的语义,导致模型不能很好地分类这两种类别. 而第二种类型的错误,我们认为主要的原因是由数据集中导致的,其中数据集中负例类别数量为 28 509,而正例数量仅有 4999 条,这不可避免地使得数量少的类别被错误地分类到数量大的实例中.

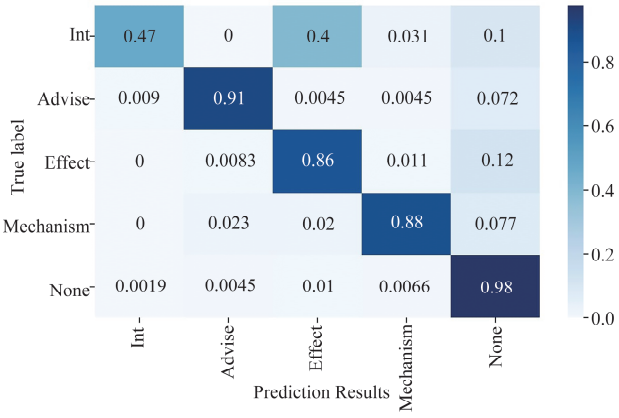

图 6　本文模型的混淆矩阵

Fig. 6　Confusion matrix of model

6　结　论

本文提出了 RBio-VGCN 模型用于 DDI 关系抽取. 该模型充分利用了 BioBERT 动态捕获输入句子和目标药物实体的上下文信息,同时基于数据集构建词汇图,与输入句子进行图卷积操作获得与句子相关的全局特征信息,并使用自注意力机制最大化获取与关系抽取任务相关的特征信息,摒弃不相关的特征信息. 实验结果表明,本文模型在 DDIExtraction 2013 关系抽取任务中取得了很好的效果. 在未来的工作中,我们会针对数据集中负例较多的数据不平衡现象,考虑数据增强等方案来平衡数据,使模型的实验性能提高.

参考文献:

[1] Segura-Bedmar I, Martinez P, Sánchez-Cisneros D. The 1st DDIExtraction-2011 challenge task: extraction of Drug-Drug interactions from biomedical texts [C]//Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction. Huelva: [s. n.], 2011.

[2] Segura B I, Martínez P, Herero Z M. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIextraction2013) [C]. [S. l.]: Association for Computational Linguistics, 2013.

[3] Kim S, Liu H, Yeganova L, *et al*. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach [J]. J Biomed Inform, 2015, 55: 23.

[4] Lu Z, Du P, Nie J Y. VGCN-BERT: augmenting BERT with graph embedding for text classification [C]//Procedings of the European Conference on Information Retrieval. Berlin: Springer, 2020.

[5] Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36: 1234.

[6] Kipf T N, Welling M. Semi-Supervised classification with graph convolutional networks[C]//Procedings of the 5th International Conference Learning Representations (ICLR). Toulon, France: ICLR, 2017.

[7] Thomas P, Neves M, Rocktäschel T, *et al*. WBI-DDI: drug-drug interaction extraction using majority voting [C]//Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Atlanta, Georgia, USA: SemEval, 2013.

[8] Zheng W, Lin H, Zhao Z, *et al*. A graph kernel based on context vectors for extracting drug - drug interactions [J]. J Biomed Inform, 2016, 61: 34.

[9] Liu S, Chen K, Chen Q, *et al*. Dependency-based convolutional neural network for drug-drug interaction extraction [C]//Proceedings of the 2016 IEEE international conference on bioinformatics and biomedicine (BIBM). Shenzhen, China: IEEE, 2016.

[10] Zhao Z, Yang Z, Luo L, *et al*. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network[J]. Bioinformatics, 2016, 32: 3444.

[11] 刘宁宁，琚生根，熊熙，等. 基于胶囊网络的药物相互作用关系抽取方法[J]. 中文信息学报，2020，34：80.

[12] Park C, Park J, Park S. AGCN: attention-based graph convolutional networks for drug-drug interaction extraction [J]. Expert Syst App, 2020, 159: 113538.

[13] Devlin J, Chang M W, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for Language Understanding [C]//Proceedings of NAACL-HLT. Minneapolis, Minnesota: NAACL, 2019.

[14] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need [C]//Proceedings of 31st Conference on Neural Information Processing Systems. Long Beach, CA, USA: NIPS, 2017.

[15] Nguyen D P, Ho T B. Drug-drug interaction extraction from biomedical texts via relation bert[C]//Proceedings of the 2020 RIVF International Conference on Computing and Communication Technologies (RIVF). Ho Chi Minh City, Vietnam: IEEE, 2020.

[16] Wu S, He Y. Enriching pre-trained language model with entity information for relation classification [C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China: CIKM, 2019.

[17] Zhu Y, Li L, Lu H, *et al*. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions [J]. J Biomed Inform, 2020, 106: 103451.

[18] Quan C Q, Hua L, Sun X, *et al*. Multichannel convolutional neural network for biological relation extraction [J]. Biomed Res Int, 2016, 2016: 1.

[19] Asada M, Miwa M, Sasaki Y. Enhancing drug-drug interaction extraction from texts by molecular structure information[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg, PA, USA: ACL, 2018.

[20] Sun X, Dong K, Ma L, *et al*. Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss [J]. Entropy, 2019, 21: 37.

[21] Zhou D, Miao L, He Y. Position-aware deep multi-task learning for drug-drug interaction extraction [J]. Artif Intell Med, 2018, 87: 1.

[22] Zheng W, Lin H, Luo L, *et al*. An attention-based effective neural model for drug-drug interactions extraction [J]. Bmc Bioinformatics, 2017, 18: 1.

[23] Kumar S S, Ashish A. Drug-Drug interaction extraction from biomedical text using long short term memory network [J]. J Biomed Inform, 2017, 86: 15.

引用本文格式：

中　文：杨霞，韩春燕，琚生根. 基于实体信息和图神经网络的药物相互作用关系抽取 [J]. 四川大学学报：自然科学版，2022，59：022002.

英　文：Yang X, Han C Y, Ju S G. Drug-Drug relationship extraction based on entity information and graph neural networks [J]. J Sichuan Univ: Nat Sci Ed, 2022, 59: 022002.