基于外部知识的药物相互作用关系抽取方法

王　芳[1]，龙　欣[2]，周　刚[2]，刘宁宁[2]
(1. 四川省医学科学院四川省人民医院辅助生殖医学中心，成都 610110；2. 四川大学计算机学院，成都 610065)

摘　要：药物相互作用是指药物之间存在的抑制或促进等作用. 针对目前方法在不同关系类别上的抽取结果差异较大的问题，提出了一种利用外部知识的关系抽取模型，该方法首先对外部药物数据库中的信息进行处理，构建带有药物描述信息的数据集；然后，在该数据集上进行模型训练，并保存最优模型；最后，将该最优模型与药物关系抽取模型相结合，进行药物关系抽取，从而更好的利用了药物数据库中已有的知识，缓解了不同关系类别抽取结果差异较大的问题，提高了抽取效果. 在 DDIExtraction2013 数据集上的实验结果表明，论文方法的 F_1 值优于目前最优方法 2.47%.

关键词：药物关系抽取；外部知识；双向长短期记忆网络；胶囊网络
中图分类号：TP391　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2021.062003

An extraction method for drug-drug relationships based on external knowledge

WANG Fang[1], *LONG Xin*[2], *ZHOU Gang*[2], *LIU Ning-Ning*[2]
(1. The Department of Assisted Reproductive Center in Sichuan Academy Medical Sciences & Sichuan Provincial People's Hospital, Chengdu 610110, China;
2. College of Compute Science, Sichuan University, Chengdu 610065, China)

Abstract: Drug-Drug Interaction refers to the inhibition or promotion between drugs. In order to solve the problem that the extraction results of the current model are quite different in different relationship categories, this paper proposes a relationship extraction model based on external knowledge. This method first processes the information in the external drug database, constructs a data set with drug description information on which the model is trained and then the optimal model is saved. Finally, the resulting optimal model is combined with the drug relationship extraction model to extract drug relationships, which can make better use of the existing knowledge in drug database, alleviate the problem of large differences in the extraction results of different relationship categories, and improve the extraction result. The experimental results on the DDIExtraction 2013 dataset show that the F1 value of the proposed approach in this paper is 2.47% higher than the current best approaches.

Keywords: Drug relationship extraction; External knowledge; Bidirectional long short term memory; Capsule network

收稿日期：2021-06-28
基金项目：四川省新一代人工智能重大专项(2018GZDZX0039)；四川省重点研发项目(2019YFG0521)
作者简介：王芳(1990—)，女，四川内江人，研究领域为辅助生殖医学. E-mail：augustwp@163.com
通讯作者：周刚. E-mail：zhougang@scu.edu.cn

1 引　言

药物和药物的相互作用(Drug-Drug Interactions,DDI)指的是多种药物被病人在同一时间服用时,不同药物之间产生的协同或拮抗等作用[1].这些作用所带来的副作用会导致治疗费用增加,严重的话,甚至会对患者的身体健康造成极大的威胁.因此了解药物和药物之间的相互作用知识对于患者的诊治和医学的发展有着非常重要的意义与价值.现阶段已经提出的药物相互作用关系提取研究,在系统生物学、个性化医学等[2]许多生物医学领域具有重要的价值.

现阶段的学者们通常从 DrugBank[3],PharmGKB[4]等医学数据库中获取药物相关知识.然而,调研发现,医学数据库只存储了少量 DDI.相比而言,相关医学文献中记录的药物之间的相互作用信息更加丰富.且现有医学数据库主要采用手动的方式更新数据库,即依靠医学专业人员从文献中手动抽取出 DDI,这种方法需要的成本高且效率较低.因此,医学数据库的更新速度远远落后于文献的增长速度,导致无法充分利用生物医学文献中海量的医学资源.药物相互作用关系抽取旨在自动从海量的医学文献中高效、准确地抽取出关系来更新医学数据库.此外,药物相互作用知识在药物研发、辅助医生开药、构建医学知识图谱、更新医学数据库、预防药物的不良反应等应用中也发挥着巨大的作用.自 2011 年和 2013 年先后发布的两个(DDIExtraction2011[5],DDIExtraction2013[6])对药物相互作用进行关系抽取的任务起,如何从文本中自动提取 DDI 逐渐成为研究热点.目前在 DDI 抽取任务上,已经有许多成功应用的方法.

早期对于 DDI 的抽取通常使用基于规则的方法,其中规则的制定一般需要医学领域中专业人员的辅助.这种方法的召回率偏低,这是因为语言表达形式具有多样性,部分药物之间相互作用关系可能不能被制定的规则所覆盖.随着机器学习技术的日益进步,药物相互作用关系开始更多地利用机器学习方法来进行抽取.这类方法的抽取性能很大一部分取决于外部自然语言处理工具的效果.这是因为该方法需要用到大量由词性标注器、句法分析器等自然语言处理工具生成的如词性,句法,语法等人工定义的特征.为了减少人工设计特征所耗费的成本且取得比传统方法更好的抽取效果,通过利用深度学习技术来自动学习特征并抽取 DDI 逐渐成为趋势.

本文提出了一种结合外部知识的药物关系抽取模型,针对药物相互作用数据集中存在的样本数目较少以及不同关系类别样本数目差异较大的问题,提出将外部药物知识融入关系抽取模型中.首先需要对 DrugBank 药物数据库中存在的药物知识进行抽取,从中抽取出带有药物描述信息的药物相互作用对,以此构建带有药物描述信息的数据集.然后通过在此数据集上进行训练,得到训练好的药物描述信息模型.最后将该模型与基于注意力机制的药物相互作用关系抽取模型相结合,即为融入外部知识的药物相互作用关系抽取模型.该模型可以有效利用药物数据库 DrugBank 中已有的知识,从而提高模型的抽取效果、缓解不同关系类别之间抽取结果差异过大的问题.最后,本文模型的有效性也在数据集 DDIExtraction2013 上得到了验证.

本文的章节结构安排如下:第 2 节介绍关于药物相互作用关系抽取的相关工作;第 3 节中详细地描述并展示了本文提出的模型;第 4 节中通过利用 DDIExtraction2013 数据集对本文模型的有效性进行了综合验证,并与其他模型进行对比;第 5 节总结全文并提出了未来的发展方向.

2 相关工作

目前,该领域主要有三类方法被应用于提取药物和药物的相互作用关系,即基于规则的方法,基于传统机器学习的方法和基于深度学习的方法.

基于规则方法的重点在于如何从医学文本中找到合适的规则.如 Segura-Bedmar 等[7]提出了一种混合方法来从生物医学文本中提取药物之间的关系,该方法结合了浅层分析和模式匹配,通过药剂师根据专业经验以及对语料库中描述 DDI 语法结构的分析观察,制定出一组特定的 DDI 抽取规则,最后在生成的简单句上应用该规则进行关系抽取.Blasco 等[8]认为在医学文本中存在特定的描述 DDI 的语句形式,故采用 Apriori 算法来提取医学文本中最大频繁序列,然后利用该序列进行 DDI 的抽取.Santiago 等[9]通过使用临床资料、科学文献和社交媒体的数据挖掘研究检测药物相互作用.数据挖掘在 DDI 分析中具有重要应用:发现药物之间的不利影响,建立知识数据库、黄金标准以及抽取新的 DDI.

基于传统机器学习方法的重心在于对各种特

征的使用以及核函数的设计. 如 Chowdhury 等[10]通过从文本中提取出一组触发词特征、否定词特征、句法特征和词汇特征，然后使用支持向量机完成 DDI 的抽取. Zheng 等[11]提出了一种新型的基于图的内核函数，该内核函数不仅计算顶点本身的属性，还计算了顶点的相邻属性，从而可以充分捕获语句的结构信息，提高了抽取效果. Zhang 等[12]针对句子结构中存在的噪声问题，提出了一种图修剪方法，可以从原始句子结构中修剪明显的噪声信息，并强调相关的句法信息. 该方法可以更准确的计算和表示语法结构信息，在当时取得了最佳结果. Chowdhury 等[13]提出了一种复合内核的方法，可以充分利用依赖特征、上下文特征以及全局特征等特征，效果比单个的核函数好.

近年来，利用深度神经网络来对药物关系进行抽取取得了较好的发展. Liu 等[14]通过将单词嵌入和位置嵌入信息输入到卷积网络中，首次提出利用 CNN 模型来进行药物关系抽取. 实验结果表明，该方法即使在不提取任何词性，句法等特征的基础上，也能取得较好的结果，不仅证明在解决药物关系抽取问题上深度学习方法是适用的，而且显示该方法相对于基规则和传统机器学习方法能取得更好的结果. 为了更好地对句子语义的表示进行获取，Kavuluru 等[15]考虑到相比词级别的嵌入，字符级别的嵌入更适用于表示形态较为丰富的自然语言，首次将字符级别的循环神经网络应用到 DDI 抽取上. Zhou 等[16]认为判断药物之间的关系，单词的位置信息有非常重要的作用，为了能更好地利用位置信息，他们提出将 BiLSTM 层产生的隐层状态与位置嵌入结合，从而生成位置感知注意力. Zhao 等[17]首次使用图卷积网络编码语法图进行医学关系的提取，通过将句法知识建模为图卷积网络中的边，将单词作为节点，利用双向门控循环单元网络学习句子的序列特征，图卷积网络学习句法图表示的特征. 实验结果表明，双向门控循环单元网络和图卷积网络的结合能够进一步提高模型性能. Asada 等[18]采用卷积神经网络和图卷积神经网络相结合的方法进行 DDI 抽取，该模型首先利用卷积神经网络对文本语句进行编码，然后利用图卷积神经网络对药物的分子结构信息进行编码，然后将二者的池化层相结合，生成最终的预测结果. 该方法能够有效利用药物分子结构信息. Zhang 等[19]通过将卷积神经网络与循环神经网络相结合，取得了比单个模型更好的结果. Feng 等[20]用图卷积网络和深度神经网络结合，提出了一种有效且鲁棒的方法，通过利用 DDI 网络信息而不是考虑药物特性来预测潜在的 DDI. 该方法在其他与 DDI 相关的场景中也有用，例如药物组合指导、检测药物副作用等.

3　本文方法

本文提出了一种结合外部知识的药物关系抽取模型，模型整体流程图如图 1 所示.

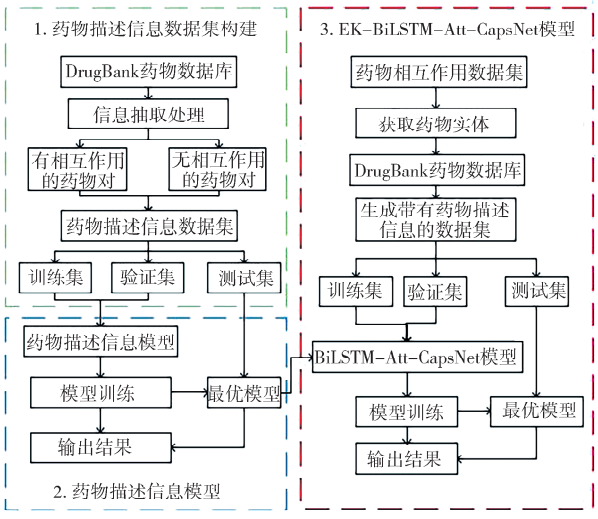


图 1　模型整体流程图

Fig. 1　Flow chart of model architecture

图 1 主要包括三个部分，分别是药物描述信息数据集构建、药物描述信息模型以及 EK-BiLSTM-Att-CapsNet 模型. 各部分的主要内容如下.

第 1 部分，药物描述信息数据集构建. 为了更好地利用外部知识来提高模型的抽取效果，在该部分中通过对药物数据库 DrugBank 中的知识进行分析和处理，从中抽取出有相互作用的药物对，并生成无相互作用的药物对，同时保留每个药物的描述信息，以此构建带有药物描述信息的药物相互作用数据集.

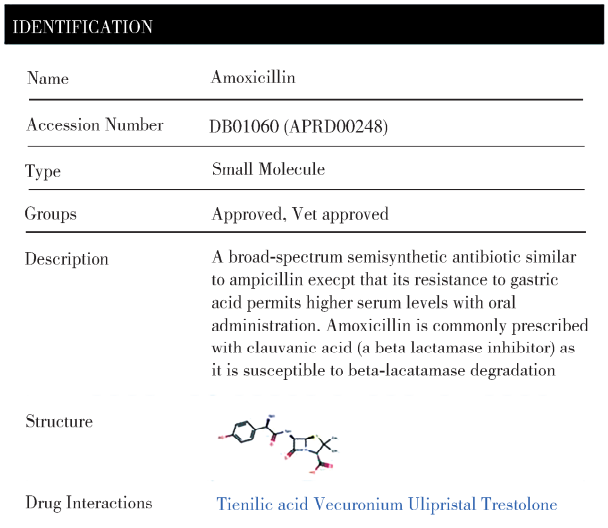
第 2 部分，药物描述信息模型. 为了更好地利用药物数据库中已有的药物相互作用知识以及药物描述的相关信息，在本部分中构建基于 BiLSTM 的药物描述信息训练模型，并在第一部分构建的药物描述信息数据集上进行训练，保存最优模型.

第 3 部分，EK-BiLSTM-Att-CapsNet 模型. 在本部分中，首先需要对药物相互作用数据集 DDIExtraction2013 进行处理，从中识别出药物实

体,再在药物数据库 DrugBank 中找到对应的药物描述信息,并将该信息保存. 然后将第二部分中保存的最优模型与 BiLSTM-Att-CapsNet 模型相结合,即为结合外部知识的药物相互作用关系抽取模型 EK-BiLSTM-Att-CapsNet. 最后通过对该模型进行训练,得到最优的模型,用于药物相互作用关系抽取.

3.1　药物描述信息数据集构建

DrugBank 是较为知名的药物数据库,通过该数据库可以查询到药物的各种信息,如药物类型、药物结构、药物描述信息、药物别名以及与该药物之间有相互作用的其他药物等相关信息. 例如在 DrugBank 数据库的查询框中输入"Amoxicillin"(阿莫西林)时,页面就会显示出所有与 Amoxicillin 相关的信息,如图 2 所示.


IDENTIFICATION

Name	Amoxicillin
Accession Number	DB01060 (APRD00248)
Type	Small Molecule
Groups	Approved, Vet approved
Description	A broad-spectrum semisynthetic antibiotic similar to ampicillin execpt that its resistance to gastric acid permits higher serum levels with oral administration. Amoxicillin is commonly prescribed with clauvanic acid (a beta lactamase inhibitor) as it is susceptible to beta-lacatamase degradation
Structure	
Drug Interactions	Tienilic acid Vecuronium Ulipristal Trestolone

图 2　DrugBank 数据库药物查询示例

Fig. 2　Example of DrugBank database inquiry

通过对药物数据库中药物的各种信息进行分析,我们发现药物的描述信息是对该药物的详细介绍. 在描述信息中一般包括该药物适用的疾病类型、组成成分、作用效果等信息. 两个药物的描述信息中,有可能隐含着这两个药物之间的潜在关系,有助于判断二者是否存在相互作用关系. 因此本文从药物的各种信息中,选用药物的描述信息作为本文模型使用的外部知识之一. 在对 DrugBank 中的信息进行处理后,从中共计获取到 11 930 个药物以及对应的描述信息.

在药物数据库 DrugBank 中,每个药物的相关信息中包含与该药物有相互作用的药物信息,通过对该信息进行处理,可以从中抽取出有相互作用的药物对,以此利用 DrugBank 中已经存在的大量的药物相互作用的相关知识. 由图 2 可以看出,与 Amoxicillin 有相互作用关系的有 Tienilic acid、Vecuronium、Ulipristal 以及 Trestolone 等药物,即可产生如下几组有相互作用的药物对:(Amoxicillin, Tienilic acid)、(Amoxicillin, Vecuronium)、(Amoxicillin, Ulipristal)、(Amoxicillin, Trestolone).

由于模型训练用到的数据集不仅需要有相互作用的药物对,还需要无相互作用的药物对,即数据集中需要正样本和负样本同时存在. 而药物数据库中不存在现成的无相互作用的药物对,故本文为了得到无相互作用的药物对,首先根据药物编号随机组合两个药物实体为一组药物对,然后从中把有相互作用的药物对过滤掉,即可得到无相互作用的药物对. 为了不对实验结果造成干扰,我们将 DDIExtraction2011 和 DDIExtraction2013 数据集中出现过的药物对过滤掉.

经过上述数据处理操作,即可构建成带有药物描述信息的药物相互作用数据集. 该数据集中的两个药物之间的关系有两种,分别是有相互作用关系和无相互作用关系. 本文通过对 DrugBank 中的药物信息进行处理,获取了总计 60 万组有相互作用的药物对,并随机生成了 60 万组无相互作用的药物对. 将其按照一定的比例划分为训练集、验证集、测试集,数据集的详细信息如表 1 所示.

表 1　药物描述信息数据集

Tab. 1　Drug description information dataset

数据集	正例/万	负例/万
训练集	56	56
验证集	2	2
测试集	2	2

3.2　药物信息描述模型

为了更好地利用药物的描述信息,我们设计了药物描述信息模型. 该模型共由五层构成,分别是输入层、嵌入层、BiLSTM 层、全连接层、输出层,如图 3 所示.

由图 3 可以看出,该模型首先将两个药物实体的描述信息作为模型的输入,并在嵌入层中将输入层的文本转换为向量表示. 然后利用 BiLSTM 层分别获取到两个药物描述信息的长序列依赖信息,并对该信息进行处理. 最后经过 softmax 等操作,得到最终的结果. 模型各层详细的内容介绍如下.

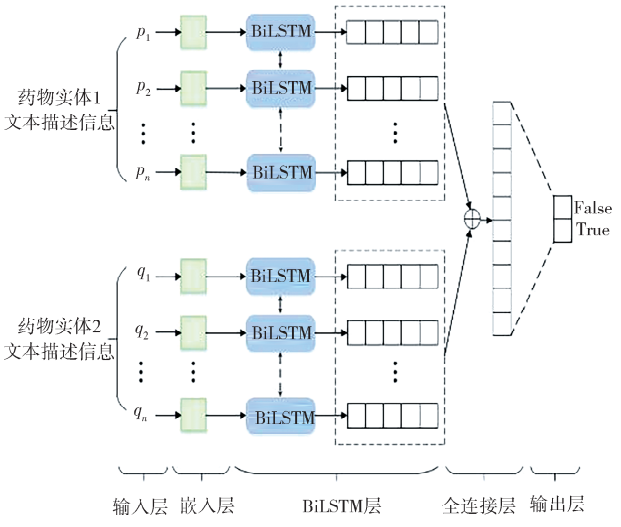


图 3　药物描述信息模型结构

Fig. 3　Structure of drug description information model

（1）输入层：在该层中，将两个药物实体对应的描述语句进行输入. 药物实体 1 的描述语句用 p 表示，$p=\{p_1,p_2,\cdots,p_i,\cdots,p_n\}$，药物实体 2 的描述语句用 q 表示，$q=(q_1,q_2,\cdots,q_i,\cdots,q_n)$. 其中 p_i 和 q_i 分别表示药物实体 1 和药物实体 2 的描述语句中的第 i 个单词.

（2）嵌入层：为了更好地利用单词的语义信息，在该层中采用 Glove 词向量表示的方式，将单词转为向量的形式表示. 经过嵌入层之后，两个药物实体的描述语句分别可以表示为 $p=(w_1^{(1)},w_2^{(1)},\cdots,w_i^{(1)},\cdots,w_n^{(1)})$，$q=(w_1^{(2)},w_2^{(2)},\cdots,w_i^{(2)},\cdots,w_n^{(2)})$. 其中 $w_1^{(1)},w_1^{(2)}\in R^{dw}$，$dw$ 表示单词嵌入的维度.

（3）BiLSTM 层：为了更好地获得输入语句的长序列依赖信息，通过使用 BiLSTM 网络，分别获取语句的前向信息和后向信息，然后将二者相结合，作为句子表示. 如式(1)和式(2)所示，其中 $\overrightarrow{S}$ 表示正向输入的语句，$\overleftarrow{S}$ 表示逆序输入的语句.

$$\overrightarrow{h}=LSTM(\overrightarrow{S}) \tag{1}$$

$$\overleftarrow{h}=LSTM(\overleftarrow{S}) \tag{2}$$

BiLSTM 隐藏层的输出可以表示为 $H=[\overrightarrow{h};\overleftarrow{h}]$. 由于 LSTM 当前时间步的输入包含了前一个时间步的信息，所以可以使用最后一个时间步的输出来表示句子. 假设 $\overrightarrow{h}_{\text{Last}}$ 表示前向输入的最后一个时间步的信息，$\overleftarrow{h}_{\text{Last}}$ 表示后向输入的最后一个时间步的信息，则经过 BiLSTM 层之后，语句可以表示为 $h=[\overrightarrow{h}_{\text{Last}};\overleftarrow{h}_{\text{Last}}]$.

将药物实体 1 的描述信息 p 经过 BiLSTM 层之后，可以表示为 $h_1\in R^N$，药物实体 2 的描述信息 q 经过 BiLSTM 层之后，可以表示为 $h_2\in R^N$. 将二者拼接，即 $h^*=[h_1;h_2]\in R^{2N}$ 送入全连接层中，其中 N 表示 BiLSTM 隐藏层单元数目.

（4）全连接层：由于 BiLSTM 的输出维度为 $2N$，而输出层的节点为关系类别数目 m，故需要使用全连接层进行线性变换. 如式(3)所示，其中 $W^{(fc)}$ 和 $b^{(fc)}$ 为全连接层的参数.

$$output=W^{(fc)}\cdot h^*+b^{(fc)} \tag{3}$$

为了提高模型的泛化能力，降低模型过拟合的风险，我们在全连接层应用 Dropout 机制. 其思想是在训练过程中，从神经网络中随机丢弃神经元（以及它们的连接），以此缓解神经元之间过度协同适应[21].

（5）输出层：在该层中选用 softmax 函数对全连接层的输出 $output$ 进行归一化处理，即将各个类别对应的值转换为对应的概率，所有类别概率值的总和为 1，并从中选取概率最大的作为预测的关系类别 $\hat{y}$，如式(4)所示.

$$\hat{y}=\text{argmax}(\text{softmax}(output)) \tag{4}$$

该模型使用的损失函数如式(5)所示. 其中，$y\in R^m$ 表示真实的类别标签；$\hat{y}\in R^m$ 表示预测的类别标签；m 表示类别标签数目；y 和 $\hat{y}$ 以 one-hot 向量表示；λ 是 L_2 正则化的超参数；θ 需要在模型中进行训练.

$$J(\theta)=-\frac{1}{m}\sum_{i=1}^{m}y_i\log(\hat{y}_i)+\lambda\|\theta\|^2 \tag{5}$$

该模型使用 3.1 节构建的带有药物描述信息的药物相互用数据集进行训练，并将取得最优结果的模型进行保存.

3.3　EK-BiLSTM-Att-CapsNet 模型

为了充分利用外部药物描述信息和药物相互作用知识，本文将 3.2 节中保存的最优的药物描述信息模型与基于注意力的药物关系抽取模型相结合，构成 EK-BiLSTM-Att-CapsNet 模型. 该模型将药物的描述信息作为低层胶囊网络的一部分，然后将其动态的传输到高层胶囊网络中，从而更好地使用药物描述信息，此外，药物描述信息模型是在构建的药物相互作用数据集上进行训练得到，该数据集是对药物数据库中已有的信息进行处理得到，故可以充分利用外部已有的药物相互作用知识. EK-BiLSTM-Att-CapsNet 的模型结构如图 4 所示.

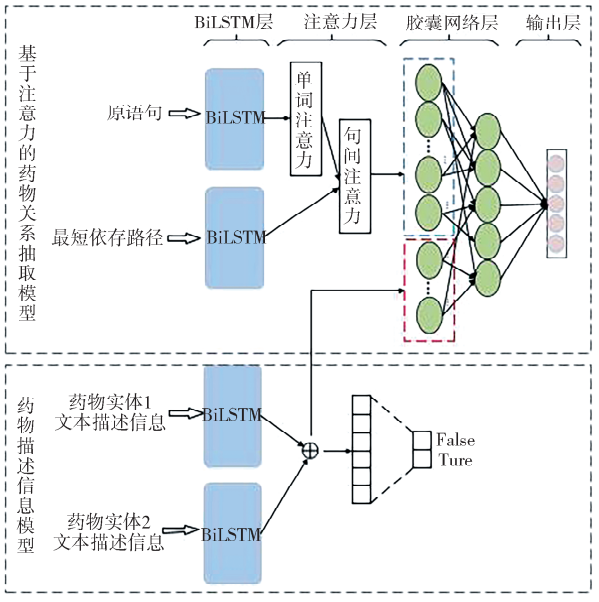


图 4　EK-BiLSTM-Att-CapsNet 模型结构

Fig. 4　Structure of EK-BiLSTM-Att-CapsNet model

假设输入原语句$S_{original}$，该语句中两个药物实体之间的最短依存路径为S_{sdp}，原语句中的两个药物实体在 DrugBank 中的描述信息分别des_1用和des_2表示.

如图 4 上部所示，将以文本表示的$S_{original}$和S_{sdp}经过嵌入层，可得到以向量形式进行表示的$S_{original}$和S_{sdp}，再将二者分别送入 BiLSTM 层，即可得到原语句的长序列依赖信息$H_{original}$和最短依存路径的长序列依赖信息表示H_{sdp}. 语句中不同单词的重要性不同，因此将单词级别的注意力机制应用在原语句$H_{original}$上，即可得到加权后的$H_{original}$. 为了缓解使用最短依存路径可能造成的噪声干扰问题，在$H_{original}$和H_{sdp}应用句子级别的注意力机制，可以有效地将原语句信息和最短依存路径信息相融合，用H_{all}表示.

如图 4 下部所示，两个药物的描述信息des_1和des_2经过嵌入层之后，我们将其分别送入 BiLSTM 层，即可得到相对应的 BiLSTM 隐藏层输出H_{des_1}和H_{des_2}，将二者结合，即可得到$H_{des}=[H_{des_1};H_{des_2}]$.

将H_{all}以及H_{des}分别经过卷积操作，得到低层胶囊表示，$u_{all}=(u_1,u_2,\cdots,u_m)$和$u_{des}=(u_1,u_2,\cdots,u_n)$，其中 m 和 n 分别表示H_{all}产生的低层胶囊个数，以及药物描述信息H_{des}产生的胶囊个数. 二者共同构成胶囊网络层的低层胶囊 $u=[u_{all};u_{des}]$. 低层胶囊向高层胶囊传输的信息量可以利用胶囊网络层的动态路由机制来动态地控制，这样就可以动态地利用药物描述信息这一外部知识.

本模型采用的损失函数，如式(6)所示.

$$L_k=T_k\max(0,m^+-\|v_k\|)^2+\lambda(1-T_k)\max(0,\|v_k\|-m^-)^2 \tag{6}$$

4　实　验

4.1　数据集介绍以及评价指标

本文采用 DDIExtraction2013 数据集来进行实验，该数据集是由 792 篇 DrugBank 中的医学文本和 MedLine 中的 233 篇摘要组成，并事先对药物实体进行了标注. 因此，识别出药物实体不需要经过命名实体识别等操作. 该数据集中共有如下 5 种药物相互作用类型.

(1) Advice：文本中描述了同时使用两种药物时的建议.

(2) Effect：文本中清楚地表明了两种药物相互作用的结果.

(3) Mechanism：文本中明确讨论了药物动力学机制.

(4) Int：文本中表明两种药物有一定关系，但具体关系类型没有被定义

(5) Negative：相互作用关系并没有存在于两种药物之间.

数据集的详细信息见表 2. 本文采用精准率，召回率和F_1值这三个指标来对实验结果进行评价.

表 2　DDIExtraction2013 数据描述

Tab. 2　DDIExtraction2013 data description

药物种别	训练集	测试集	总数
Advice	826	221	1 047
Effect	1 687	360	2 047
Mechanism	1 319	302	1 621
Int	188	96	284
Negative	23 772	4 737	28 509

4.2　参数设置

我们将药物描述信息的长度设置为 50，并将单词转换为 300 维的 Glove 词向量. 对于未在 Glove 词表中出现的单词，我们采用随机初始化的方式表示该单词的词向量. 胶囊网络的迭代次数设置为 3，Batchsize 设置为 64，dropout 取值为 0.5，学习率设置为 0.001，药物描述信息模型的迭代次数设置为 50，EK-BiLSTM-Att-CapsNet 模型的迭代次数设置为 35，模型损失函数中的m^+为 0.9，m^-为 0.1，λ为 0.25.

4.3　实验结果

为了验证模型的有效性，我们从以下三个方面进行实验，分别是消融实验、不同关系类别上的结果对比以及与现有方法的对比实验.

（1）消融实验. 为了验证本章模型的有效性，即验证本章模型中利用的药物描述信息以及药物数据库 DrugBank 中已有的相互作用知识是否有用，设计消融实验，在 DDIExtraction2013 数据集上的实验结果，如表 3 所示.

表 3　消融实验结果

Tab. 3　Results of ablation experiments

模型	P/%	R/%	F_1/%
BiLSTM-Att-CapsNet	79.69	70.35	74.73
＋药物描述信息	79.47	71.01	75.00
＋药物描述信息＋DrugBank 中 DDIs 信息	80.11	71.32	75.46

由消融实验结果可以看出，在直接添加外部的药物描述信息后，模型的 F_1 值提高了 0.27%，说明药物描述信息中存在对药物相互作用关系判断的信息. 通过使用该信息，可以提高模型的抽取效果. 在使用 DrugBank 中的带有药物描述信息的药物相互作用数据集进行训练后，模型的效果提升了 0.73%. 这证明了外部药物数据库中已有的知识对于模型抽取效果的提升有很大的帮助. 消融实验的结果表明，我们提出的模型可以有效利用外部已有的药物描述信息，同时可以利用 DrugBank 中已有的大量的药物相互作用知识，有助于提高模型的抽取结果.

（2）不同关系类别上的结果对比. 通过对数据集的分析可知，数据集中不同类别的数目相差较大，因此可能造成不同类别的预测结果相差较大. 为了验证利用药物描述信息能够缓解该问题，将本文模型与相关模型在不同类别上的实验结果进行对比，具体信息如表 4 所示. 其中 Xu 等[22]提出的 BR-LSTM 和 Sahu 等[23]提出的 Joint AB-LSTM 都是基于 LSTM 的模型.

由表 4 可以看出，对比模型的不同类别之间的最大 F_1 值差值在 31.35%到 36.15%之间浮动，与他人模型相比，使用本文模型不同类别之间的 F_1 值最大相差仅为 25.34%，较大降低了不同类别的 F_1 值之间的差距. 由此可见，与其他模型相比，本章提出的模型能够有效缓解因不同类型样本数目差异较大造成的抽取结果差异较大的问题. 更直观的结果如图 5 所示.

表 4　不同关系类别上的实验结果

Tab. 4　Experimental results on different relationship categories

模型名称	F_1 值/%				差值/%
	advice	Mechanism	effect	int	
CNN[14]	77.72	70.23	69.32	46.37	31.35
PM-BLSTM[16]	81.6	74.42	71.28	48.57	33.03
CNN-GCN[18]	81.62	73.83	71.03	45.83	35.79
BR-LSTM[22]	75.18	79.11	68.17	43.36	35.75
Joint AB-LSTM[23]	80.26	72.26	65.46	44.11	36.15
本文模型	77.38	80.41	74.08	55.07	25.34

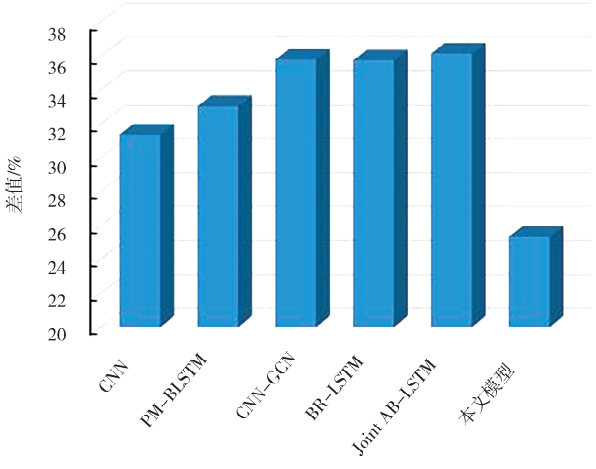


图 5　不同类别间的最大 F_1 差值

Fig. 5　Maximum F_1 score difference between different categories

由图 5 可以明显看出，与其他模型相比，本章提出的 EK-BiLSTM-Att-CapsNet 模型的不同类别之间的 F_1 值差距较小，能够较好地缓解数据集中因不同类别的数目差异较大造成的不同类别的 F_1 值差异较大的问题.

（3）与现有方法的对比实验. 目前在药物相互作用关系抽取任务上，已有大量的研究人员设计出了多种关系抽取模型，并进行了大量实验. 为了验证本文模型的有效性，将本文模型与现有模型在 DDIExtraction2013 数据集上的实验结果进行对比.

本文对比了不同模型在 DDIExtraction 2013 数据集上的实验结果. 其中 Quan 等[24]提出的 MCCNN 基于卷积神经网络进行自动特征提取；Wang 等[25]的模型通过将基于依赖的技术引入 Bi-LSTM，建模药物之间的依赖关系；Yi 等[26]的模型

通过将一个循环神经网络结合多个注意力层从生物医学中提取 DDI 文本. 实验结果如表 5 所示.

表 5　不同模型在 DDIExtraction2013 数据集上的实验结果

Tab. 5　Experimental results of different models on the DDIExtraction2013 dataset

数据来源	P/%	R/%	F_1/%
文献[14]	75.72	64.66	69.75
文献[24]	75.99	62.65	70.21
文献[25]	72.53	71.49	72.00
文献[26]	73.67	70.79	72.20
文献[22]	71.52	70.79	71.15
文献[18]	73.31	71.81	72.55
文献[16]	75.8	70.38	72.99
本文	80.11	71.32	75.46

由表 5 可以看出,本文模型的 F_1 值比其他文献中最佳模型高 2.47%. 实验结果表明,与现有模型相比[27-28],本文提出的模型能够更好地自动抽取药物之间的相互作用关系. 更直观的结果如图 6 所示.

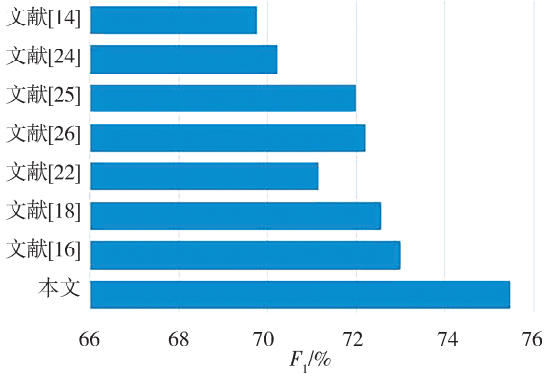


图 6　不同模型在 DDIExtraction2013 数据集上的 F_1 值

Fig. 6　F_1 score of different models on the DDIExtraction2013 dataset

5　结　论

本文充分利用了外部药物数据库中的相关信息,使用胶囊网络的结构进行药物关系抽取. 实验结果表明,本文模型不仅提升了抽取效果,还降低了不同类别抽取结果差异较大的问题. 在未来的工作中可以尝试借助药物的其他信息进行辅助. 此外,还可以考虑使用远程监督的方法,用于解决数据集中样本数目过少以及不同类型样本数目差异较大的问题.

参考文献:

[1] Baxter K. Stockley's drug interactions [J]. Int J Clin Pract, 2010, 58: 226.

[2] Sänger M, Leser U. Large-scale entity representation learning for biomedical relationship extraction [J]. Bioinformatics, 2021, 37: 236.

[3] Wishart D S, Feunang Y D, Guo A C, *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018 [J]. Nucleic Acids Res, 2018, 46: D1074.

[4] Thorn C F, Klein T E, Altman R B. PharmGKB: the pharmacogenomics knowledge base [M]//Pharmacogenomics. Totowa: Humana Press, 2013.

[5] Segura-Bedmar I, Martínez P, Sánchez-Cisneros D. The 1st DIExtraction-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.

[6] Segura-Bedmar I, Martínez P, Herrero Z M. Semeval-2013 task 9: extraction of drug-drug interactions from biomedical texts (ddiextraction 2013) [C]. [S. l.]: Association for Computational Linguistics, 2013.

[7] Segura-Bedmar I, Martínez P, de Pablo-Sánchez C. A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents [J]. BMC Bioinformatics, 2011, 12: S1.

[8] Blasco S G, Danger R, Rosso P. Drug-drug interaction detection: a new approach based on maximal frequent sequences [J]. Procesamiento Del Lenguaje Natural, 2010(45): 263.

[9] Santiago V, Carol F, George H. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media [J]. Brief Bioinform, 2017(5): 5.

[10] Chowdhury M F M, Abacha A B, Lavelli A, *et al*. Two different machine learning techniques for drug-drug interaction extraction [C]// Proceedings of DDIExtraction2011: First Challenge Task: Drug-Drug Interaction Extraction. Huelva: [s. n.], 2011.

[11] 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.

[12] Zhang Y, Lin H, Yang Z, *et al*. A single kernel-based approach to extract drug-drug interactions from biomedical literature [J]. PLoS One, 2012, 7: e48901.

[13] Chowdhury M F M, Lavelli A. Drug-drug interaction extraction using composite kernels [C]// Pro-

ceedings of DDIExtraction2011：First Challenge Task：Drug-Drug Interaction Extraction. Huelva：[s. n.]，2011.

[14] Liu S Y，Tang B Z，Chen Q C，*et al*. Drug-drug interaction extraction via convolutional neural networks [J]. Comput Math Method M，2016，2016：1.

[15] Kavuluru R，Rios A，Tran T. Extracting drug-drug interactions with word and character-level recurrent neural networks [C]// Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI). [S. l.]：IEEE，2017.

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

[17] Zhao D，Wang J，Lin H，*et al*. Extracting drug-drug interactions with hybrid bidirectional gated recurrent unit and graph convolutional network [J]. J Biomed Inform，2019，99：103295.

[18] Asada M，Miwa M，Sasaki Y. Enhancing drug-drug interaction extraction from texts by molecular structure information [J]. ACL，2018(2)：680.

[19] Zhang Y，Lin H，Yang Z，*et al*. A hybrid model based on neural networks for biomedical relation extraction [J]. J Biomed Inform，2018，81：83.

[20] Feng Y H，Zhang S W，Shi J Y. DPDDI：a deep predictor for drug-drug interactions [J]. BMC Bioinformatics，2020，21：1.

[21] Srivastava N，Hinton G，Krizhevsky A，*et al*. Dropout：a simple way to prevent neural networks from overfitting [J]. J Mach Learn Res，2014，15：1929.

[22] Xu B，Shi X，Zhao Z，*et al*. Leveraging biomedical resources in bi-lstm for drug-drug interaction extraction [J]. IEEE Access，2018，6：33432.

[23] Sahu S K，Anand A. Drug-drug interaction extraction from biomedical texts using longshort-term memory network [J]. J Biomed Inform，2018，86：15.

[24] Quan C，Hua L，Sun X，*et al*. Multichannel convolutional neural network for biological relation extraction [J]. Bio Med Research International，2016，2016：10.

[25] Wang W，Yang X，Yang C，*et al*. Dependency-based long short term memory network for drug-drug interaction extraction [J]. BMC Bioinformatics，2017，18：578.

[26] Yi Z，Li S，Yu J，*et al*. Drug-drug interaction extraction via recurrent neural network with multiple attention layers [C]// Proceedings of the International Conference on Advanced Data Mining and Applications. Switzerland：Springer，Cham，2017.

[27] 马玉环，张瑞军，武晨，等. 深度残差网络和 LSTM 结合的图像序列表情识别[J]. 重庆邮电大学学报：自然科学版，2020，32：874.

[28] 蔡英凤，朱南楠，邰康盛，等. 基于注意力机制的车辆行为预测[J]. 江苏大学学报：自然科学版，2020，41：125.

引用本文格式：

中　文：王芳，龙欣，周刚，等. 基于外部知识的药物相互作用关系抽取方法[J]. 四川大学学报：自然科学版，2021，58：062003.

英　文：Wang F，Long X，Zhou G，*et al*. An extraction method for drug-drug relationships based on external knowledge [J]. J Sichuan Univ：Nat Sci Ed，2021，58：062003.