基于平行交互注意力网络的中文电子病历实体及关系联合抽取

李丽双*，王泽昊，秦雪洋，袁光辉
大连理工大学
计算机科学与技术学院
辽宁，大连
lils@dlut.edu.cn, dutzehao@mail.dlut.edu.cn
qinxueyang@snnu.edu.cn, 476708484@qq.com

摘要

基于电子病历构建医学知识图谱对医疗技术的发展具有重要意义，实体和关系抽取是构建知识图谱的关键技术。本文针对目前实体关系联合抽取中存在的特征交互不充分的问题，提出了一种平行交互注意力网络(PIAN)以充分挖掘实体与关系的相关性，在多个标准的医学和通用数据集上取得最优结果；当前中文医学实体及关系标注数据集较少，本文基于中文电子病历构建了实体和关系抽取数据集(CEMRIE)，与医学专家共同制定了语料标注规范，并基于所提出的模型实验得出基准结果。

关键词： 实体关系联合抽取 ； 双向特征交互模块 ； 自注意力机制 ； 中文电子病历 ； 数据集标注与构建

Parallel Interactive Attention Network for Joint Entity and Relation Extraction Based on Chinese Electronic Medical Record

LiShuang Li*, Zehao Wang, Xueyang Qin, Guanghui Yuan
School of Computer Science and Technology
Dalian University of Technology
Dalian, China
lils@dlut.edu.cn, dutzehao@mail.dlut.edu.cn
qinxueyang@snnu.edu.cn, 476708484@qq.com

Abstract

The construction of medical knowledge graph based on electronic medical records is of great significance to the development of medical technology, where entity and relation extraction plays a pivotal role. To solve the issue of insufficient interaction in the current joint entity and relation extraction approaches, we propose a Parallel Interactive Attention Network (PIAN) which can fully exploit the correlation between entity and relation, achieving the state-of-the-art results on standard datasets. Since there are few Chinese medical entity and relation annotation datasets, we construct an entity and relation extraction dataset based on Chinese electronic medical records (CEMRIE), formulate the corpus annotation specification with medical experts, and give the benchmark results based on our proposed model.

Keywords: Joint Entity and Relation Extraction , Bidirectional Feature Interaction Module , Self-Attention Mechanism , Chinese Electronic Medical Record , Dataset Annotation and Construction

基金项目：大连市科技创新基金(2020JJ26GX035)

1 引言

电子病历中记录了丰富的临床医学信息，例如疾病、症状等重要的医学实体，以及各类型医学实体之间的语义关系。随着医疗信息化的高速发展，各医院已经积累了海量的电子病历数据，但如何高效地从非结构化的电子病历文本中提取有价值的医疗信息仍是难题。目前，越来越多的研究者利用深度学习技术来自动抽取电子病历中的关键信息。其中知识图谱是一种能将信息要素结构化、规范化并以图的形式清晰展示的技术。构建基于电子病历的知识图谱可以对电子病历中的医学知识进行结构化描述，帮助医生和大众更方便的获取想要的知识，同时构建大规模知识图谱(Lan et al, 2021)也能为医疗问答、辅助决策等下游应用提供重要的技术支持。

其中，实体和关系抽取是构建知识图谱的关键技术之一，其主要目的是利用相关技术从结构化、半结构化或者自然语言中抽取得到实体关系三元组。实体和关系抽取主要分为流水线方法和联合抽取方法，目前在医学领域，实体及关系抽取主要基于流水线的方法，该方法将实体识别和关系抽取视为两个独立的任务。对于一段文本，首先使用实体识别模型抽取出所有实体，然后再利用关系抽取模型判断每个实体对的关系类别。命名实体识别作为关系抽取的研究基础和关键，二者之间联系密切。传统的基于流水线方法忽略了两个任务之间潜在的关联性，并且存在误差传播的问题，实体识别阶段的错误实体会严重影响关系抽取模型的结果。

为了解决流水线方法存在的问题，有研究考虑对实体识别和关系抽取进行联合学习。在通用领域，实体关系联合抽取主要分为统一编码和共享参数两种方法。统一编码方法(Zheng et al, 2017; Wang et al, 2020; Ren et al, 2021; Wang et al, 2021)将实体和关系编码到统一的标签空间，并学习统一的特征来同时表示实体与关系。然而，该类方法使用同一个模型对两个任务进行编码，一个任务的特征可能会与另一个任务特征产生冲突，导致特征混淆的问题，损害模型的整体性能。共享参数方法(Miwa and Bansal, 2016; Bekoulis et al, 2018; Wei et al, 2020; Yan et al, 2021) 通常采用相互独立的网络分别为实体与关系编码不同的特征表示，两个任务通过共享输入特征以及部分网络参数实现信息交互，可以避免特征混淆的问题。统一编码与共享参数各有其优势及局限性，共享参数的方法对实体和关系分别独立编码，克服了统一编码方法中的两个任务间特征冲突的问题，但两个任务间不能充分交互。

在医学领域，目前采用联合方法进行实体和关系抽取的相关研究较少，为了充分利用医学实体识别与关系抽取之间的密切联系，构建高质量的医学知识图谱，本文构建了一个平行交互注意力网络(Deep Interactive Attention Network, PIAN)，并应用于中文电子病历，进行中文医学实体和关系的联合抽取。模型采用两个平行的神经网络来分别编码医学实体和关系以抽取两种任务的特征，避免了两个任务的特征混淆问题。同时利用双向特征交互模块(Bidirectional Feature Interaction Module, BFIM)用于双向建模两个任务间深度的特征交互。具体地，BFIM可以使一个任务中每个字符的特征与来自另一任务对应的字符特征进行融合，并自适应地学习融合的比例。因此对于另一任务中有潜在价值的特征，BFIM可以提高这些特征的融合比例，来获取对当前任务有价值的信息。由于BFIM对称地建模了特征融合，并且使两个任务中每个字符的特征都参与了交互，因此其能够避免具有潜在价值的特征丢失并实现双向交互，从而实现医学实体和关系的深度融合。同时，为使实体识别与关系抽取两个任务能够更好的学习各自不同的任务特征，本文采用注意力机制(Self-Attention)(Guo et al, 2021)学习各自的任务特征表示。

目前，高质量的中文医学实体和关系联合抽取数据集较为匮乏，公开的主要有中文医学信息抽取数据集(Chinese Medical Information Extraction dataset, CMeIE)。中文电子病历中包含众多的生物医学实体以及实体间丰富的语义关系，本文基于真实的电子病历文本，首先通过分析大量中文电子病历语料的语言结构特点，与专家共同制定了实体及关系的标注规则，最终形成了一套标准的数据集标注流程与规范。然后构建了一个中文电子病历实体关系抽取数据集(Chinese Electronic Medical Record Information Extraction dataset, CEMRIE)，并利用所构建的平行交互注意力网络模型进行实体关系联合抽取，在实体和关系上的F1值分别达到了89.7%和80.4%的基准结果。

2 相关工作

2.1 流水线方法

医学领域的实体关系抽取主要是基于流水线的方法，其将实体识别与关系抽取视为两个独立的任务。对于医学实体识别，其模型大部分是基于LSTM-CRF结构(Rei et al, 2016)并

结合注意力机制(Luo et al, 2018)和预训练预训练语言模型(Lee et al, 2020)。与实体识别任务类似，当前主流的医学关系抽取模型主要采用CNN(Zeng et al, 2014)、注意力机制(Yi et al, 2017)、Transformer结构(Christopoulou et al, 2020)、预训练模型结合外部知识(Sun et al, 2020)、联邦学习(Sui et al, 2020)等结构。除此之外，图结构能够很好地建模层次结构复杂的生物医学长句，有利于关系特征的提取，因此图神经网络模型(Park et al, 2020)也被广泛应用于医学领域的关系抽取。Zheng等(Zheng et al, 2021)将实体的类别信息作为标签插入到抽取实体的两端，再通过关系抽取模型判断关系，取得了优异的效果。然而流水线方法忽略了两个任务的相关性以及依赖关系，并且其存在的错误传播问题也限制了流水线方法的性能，因此研究人员提出使用联合方法来同时抽取实体和关系。

2.2 联合抽取方法

2.2.1 通用领域的联合方法

在通用领域，当前主要有统一编码和共享参数两种方法。统一编码采用联合的标注策略，Zheng等(Zheng et al, 2017) 提出将命名实体识别和关系抽取联合抽象为一种序列标注任务，但是无法解决关系重叠问题。Wang等(Wang et al, 2020)提出单阶段抽取框架，以全新的字符对链接的角度解决了关系重叠的问题。随后Bekoulis等(Bekoulis et al, 2018)提出表填充的方法，即将实体和关系共同填入一个二维表中。由于二维表比一维序列具有更强的表示能力并且能够很好的表示嵌套实体以及重叠关系，因此越来越多的方法采用基于表填充的解码策略。Wang等(Wang et al, 2021)设计了三步走的近似解码策略同时解码出实体和关系，并采用了双仿射注意力机制来学习头实体和尾实体之间的相关性，然而，由于使用同一模型编码，两个任务的特征可能会彼此冲突，导致特征混淆的问题，从而降低模型性能。

基于共享参数的方法为两个任务分别学习独立的特征表示，避免了特征混淆，但是需要显式地建模任务间交互以利用其相关性。Miwa等(Miwa and Bansal, 2016)采取端到端神经网络模型，通过双向LSTM以及Tree-LSTM捕获单词序列信息和依存句法树结构信息。Bekoulis等(Bekoulis et al, 2018)提出了多头选择机制用于联合抽取。Wei等(Wei et al, 2020)提出了一种级联二进制标记框架，先抽取主体实体，再以关系作为条件抽取客体实体。然而这些方法仅通过共享输入或部分网络参数来实现信息共享，两个任务未能实现充分交互。Wang等(Wang and Lu, 2020)采用独立的序列编码器和表编码器来分别编码实体和关系任务，两个编码器之间存在显式交互，但是由于实体任务采用序列编码，其无法处理嵌套实体，表示能力有限。Yan 等(Yan et al, 2021)提出了分区过滤网络(Partition Filter Network)用于分别学习NER特征、RE特征以及共享特征，然而该模型在过滤阶段会裁剪掉部分特征，可能会丢失潜在有用信息。

2.2.2 医学领域的联合方法

目前，医学领域的联合抽取方法较少，Li等(Li et al, 2017) 构建了基于双向LSTM-RNN的联合学习模型，用于药物不良事件提取(Adverse Drug Extraction, ADE)，该模型采用共享参数的方式实现任务共享，但实际还是将两任务先后分开处理，仍然会产生误差传播。Lu等(Luo et al, 2020)提出一种基于标注策略的生物医学联合学习模型，将命名实体识别和关系抽取联合抽象为一种序列标注任务，通过合并两个任务的类型标签设计了一种新的标注方案和提取规则，但是无法解决实体关系重叠问题。Fei等(Fei et al, 2021)提出了基于Span的联合模型，采用双仿射注意力机制、语义依存分析以及图卷积网络，着重于解决实体关系重叠的问题，但是基于Span的方法会产生大量冗余跨度对，给模型引入噪声并且增加计算消耗。

3 模型

本文将医学实体关系联合抽取定义为$\text{Triple}\{\text{E}_1, \text{R}, \text{E}_2\} = \text{Model}(\text{S})$，其输入为句子序列S，输出为模型抽取出来的三元组Triple，其中包含两个实体E_1, E_2以及实体间的对应关系R。模型的主要结构如图1所示，实体识别和关系抽取两个任务分别采用相互平行的网络结构并结合特征交互模块。模型输入为字符序列，首先经过预训练语言模型编码，分别为实体识别和关系抽取生成初始特征向量，再通过BFIM进行特征交互，然后通过SAM学习各自任务的特征表示；两个任务通过“交互—学习—交互”的形式，逐步提高了两种特征的质量；最后将两个任务分别用表填充的方式进行解码并计算损失。

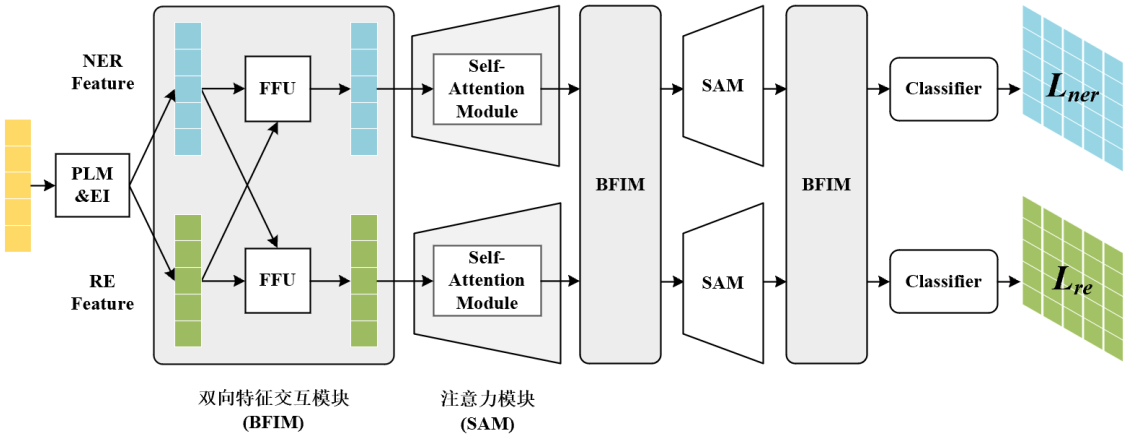

Figure 1: 平行交互注意力网络总体框架

3.1 初始特征生成

对于给定的输入句子$S = \{w_1, w_2, ..., w_n\}$，$n$表示输入序列长度，首先使用预训练语言模型例如BERT(Devlin et al, 2018)等，进行编码获得具有上下文信息的特征表示，然后分别为实体识别与关系抽取两个任务生成初始特征，考虑到实体识别和关系抽取既存在差异性又包含相关性。因此，模型采用两个独立的前馈神经网络(Feed Forward Network, FFN)并结合Dropout机制分别为两个任务生成相应的初始特征:

$$H_{bert} = \text{BERT(S)} = \{h_1, h_2, ..., h_n\}, \tag{1}$$

$$H_e = \text{Linear}_{Drop}(H_{bert}), \tag{2}$$

$$H_r = \text{Linear}_{Drop}(H_{bert}), \tag{3}$$

其中，H_e和H_r分别表示实体识别和关系抽取的初始特征表示。FFN的输入均为H_{bert}，保证了两种任务特征的相关性；并通过不同的FFN结合Dropout机制随机删除部分神经元，实现两种任务特征的差异化。

3.2 双向特征交互模块

为了建模任务间的信息交互并充分挖掘实体识别与关系抽取的相关性，本文提出了一个双向特征交互模块(BFIM)，如图1所示。该特征交互模块由两个独立的特征融合单元(Feature Fusion Unit, FFU)组成，用于融合两种任务特征，并为两个任务分别生成新的特征表示。对于来自两个任务同一位置的字符特征h_e^i, $h_r^i \in \mathcal{R}^{1\times D}$(D为字符特征维度)，首先使用一个线性层Linear将字符特征维度映射为1并拼接为$[\text{Linear}^{\text{e}}(h_e^i); \text{Linear}^{\text{e}}(h_r^i)] \in \mathcal{R}^{n\times 2}$(n为输入序列长度)，然后使用$Softmax$归一化获得融合分数(即两个任务的融合比例)，最后用融合分数与原始特征h_e^i, h_r^i做点乘操作并相加，获得某一任务的融合特征$h_{\triangle}^{*i}$:

$$(\gamma_{e,e}^i, \gamma_{e,r}^i) = Softmax([\text{Linear}^{\text{e}}(h_e^i); \text{Linear}^{\text{e}}(h_r^i)]), \tag{4}$$

$$(\gamma_{r,e}^i, \gamma_{r,r}^i) = Softmax([\text{Linear}^{\text{r}}(h_e^i); \text{Linear}^{\text{r}}(h_r^i)]), \tag{5}$$

$$\begin{bmatrix} h_e^{*i} \\ h_r^{*i} \end{bmatrix} = \begin{pmatrix} \gamma_{e,e}^i & \gamma_{e,r}^i \\ \gamma_{r,e}^i & \gamma_{r,r}^i \end{pmatrix} \begin{bmatrix} h_e^i \\ h_r^i \end{bmatrix}, \tag{6}$$

其中，Linear^{e}和Linear^{r}分别表示实体识别和关系抽取任务中FFU的线性层，$\gamma_{e,x}^i$和$\gamma_{r,x}^i$分别代表特征h_x^i在实体或关系任务新融合的特征中所占比例，[;]代表连接操作。FFU用于自适应地学习当前任务每个字符特征与另一任务对应的字符特征的最佳融合比例。

双向特征交互模块能够使每一个任务都能够自适应地融合来自另一任务的特征，同时实现更细粒度的字符级特征交互，即每个字符特征h_x^{*i}都会学习到不同的融合比例$\gamma_{x,e}^i$,$\gamma_{x,r}^i$，实现实体与关系特征之间的深度交互，从而充分利用任务间潜在有价值的特征。本文在5.4.2详细分析了每个融合单元FFU中两种任务的融合比例，验证了本文提出的BFIM的有效性。

3.3 注意力模块

注意力机制广泛应用于计算机语言以及视觉领域，并且取得了显著的效果。其中最具代表性的是自注意力机制(Self-Attention)，其可以加强重要信息的权重，并减弱干扰信息的影响。本文为实体识别与关系抽取两个任务分别设置了独立的自注意力网络，用于学习两种不同的任务特征。自注意力的计算如下，并采用“多头”模式增强注意力特征的表示能力，使模型共同处理来自不同表示子空间的信息：

$$H_{\triangle}^{multi} = [head_0; ...; head_s]W^O, \tag{7}$$

$$head_i = Attention(H_{\triangle}^{*}W_i^Q, H_{\triangle}^{*}W_i^K, H_{\triangle}^{*}W_i^V), \tag{8}$$

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

其中，W_i^Q，W_i^K，W_i^V，W^O为可训练参数，$\triangle$表示实体识别或关系抽取中的某个任务，s为注意力头数，[;]代表连接操作，经过注意力机制每个任务学习到各自具体的任务特征，之后再通过多层级的交互以及注意力网络的学习，不断提升任务特征的质量。然而随着模型深度的增加，网络可能会发生退化问题(Degradation problem)，即准确率会趋于饱和并极速下降。为了解决这一问题，在注意力模块后采用残差网络(He et al, 2016)结构并结合层归一化(Layer Normalization, LN)进行处理，得到最终输出$H_{\triangle}^{out}$：

$$H_{\triangle}^{out} = \mathrm{LN}(H_{\triangle}^{multi} + H_{\triangle}^{*}). \tag{10}$$

3.4 分类模块

本文采用基于表填充的方法对两个任务特征进行分类，将句中每个字符的特征分别与所有字符组成实体对特征，得到二维的特征表示。对于实体识别，使用实体的开始字符特征h_i与结束字符特征h_j拼接表示该实体，得到了该实体的特征表示$[h_i^e; h_j^e]$，然后通过线性层并使用ELU及$sigmoid$函数激活，得到该实体的概率分布：

$$h_{i,j}^e = \mathrm{ELU}(\mathrm{Linear}([h_i^e; h_j^e])), \tag{11}$$

$$e_{i,j}^k = \sigma(\mathrm{Linear}(h_{i,j}^e)), \tag{12}$$

与实体识别相同，关系抽取也采用基于表填充方式，使用头实体的开始字符与尾实体的开始字符来表示两个实体之间存在关系。通过线性层和激活函数得到实体对的关系特征$h_{i,j}^r$以及关系概率分布$r_{i,j}^\tau$：

$$h_{i,j}^r = \mathrm{ELU}(\mathrm{Linear}([h_i^r; h_j^r])), \tag{13}$$

$$r_{i,j}^\tau = \sigma(\mathrm{Linear}(h_{i,j}^r)). \tag{14}$$

3.5 任务训练

对于给定的训练数据集，模型的训练损失函数$\mathcal{L}$由实体识别任务的损失$\mathcal{L}_{ner}$和关系抽取任务的损失$\mathcal{L}_{re}$组成：

$$\mathcal{L}_{ner} = \Sigma_{i,j\in(1,n),\varepsilon\in\mathcal{E}}\mathrm{BCELoss}(e_{i,j}^{\varepsilon}, \bar{e}_{i,j}^{\varepsilon}), \tag{15}$$

$$\mathcal{L}_{re} = \Sigma_{i,j\in(1,n),\tau\in\mathcal{T}}\mathrm{BCELoss}(r_{i,j}^{\tau}, \bar{r}_{i,j}^{\tau}), \tag{16}$$

其中$\bar{e}_{i,j}^{\varepsilon}$与$\bar{r}_{i,j}^{\tau}$表示该实体和关系的真实标签，对两个任务均采用二分类交叉熵损失(Binary CrossEntropy Loss)函数计算各自任务损失，模型的总损失为$\mathcal{L} = \mathcal{L}_{ner} + \mathcal{L}_{re}$。

4 中文电子病历数据集构建

4.1 语料分析及预处理

本文的电子病历语料来源于某医院去隐私化后的电子病历，共包含四个科室：内科、外科、妇产科和皮肤性病科。病历中包含多个标签，例如“一般资料”、“主诉”、“现病史”、“查

体”等。每个病历中标签的类型不同，标签中的文本都包含着医疗信息，所对应的医疗信息类型区别也较大，如“查体”标签中包含较多的检查信息，“临床诊断”标签包含较多的治疗信息。对于不同科室的病历，由于医生之间的书写习惯、不同科室之间的规范不同，所包含的标签种类也不一致，这充分说明了中文电子病历领域不同样本之间的差异性，迫切需要一个统一的框架来兼容这种差异性。不仅是结构上的差异，文本语言也有很大不同，由于病历是由不同医生撰写，而医生的学历不同、籍贯不同、性别职位不同等原因会造成内容上的差异性。

除病历之间的差异性外，电子病历本身也有其特点。首先病历的组织形式属于半结构化文本，各个标签下的语言大都不是完整的一句话，语言较为简洁化、专业化；其次，病历文本中包含有大量的医学符号、数字，这些内容不同于在通用领域的含义，在医学上有其特殊意义；除此之外，病历中还包含有很多缩略词、同义词、医学专用术语等。这些病历特点给后期信息抽取带来了很大的困难。

本文将电子病历语料预处理的流程主要分为语料拆分、标签归类、标签数量统计和筛选。

语料拆分：电子病历是一种半结构化数据，包含标签及其对应的文本，对于一份完整的电子病历，本文需先根据其中的标签对病历进行拆分，每个标签下对应的为用自然语言形式描述的医学知识，方便模型处理，而且使每一段文本与对应的标签相关联，可以更精准的标识语料所包含的知识的类型。

标签归类：每个病历包含多个标签，标签表明该文本具有某种信息类型，标签之间有的差别较大，有的较为相近，统一处理会有差异性，例如标签“体格检查”和“诊疗经过”，前者包含更多的检查信息，后者包含更多的治疗信息，而“门诊化验”、“入院检查”等都与检查相关。本文先对标签进行手工归类，将包含相同方向医学知识的标签放在一个集合，后续再进行标签数量的统计和标签筛选。

标签数量统计和筛选：将标签按其所包含医学方向分类后，统计所有集合包含标签的数量，按数量多少对标签集合进行排序，从结果上看，标签“一般资料”、“主诉”、“现病史”、“既往史”数量较多，每个标签都包含着丰富的医疗信息，同时不同标签之间也会有差异性，这样分类排序之后可以更清晰的分析语料中所包含各类知识的数量比。依据统计的标签集合数量从中抽取了几个包含标签较多的集合作为本文构建数据集的语料。

4.2 数据标注及标注规则介绍

借鉴统一医学语言系统(Unified Medical Language System, UMLS)概念标准及各个中文医疗语料库的构建标准，结合抽取到的中文电子病历语料结构特点，与医学专家讨论制定了一套中文电子病历语料标注规范。

在标注规范中，本文将实体类型尽量泛化以适应语料中样本间的差异性，共定义了疾病、部位、症状、检查和治疗五种类型的实体，其详细定义如表1所示。命名实体的标注规则遵循实体间不重叠、不嵌套、实体内不含有标点符号的原则。根据确定的实体类型，本文将实体之间的关系类型分为七个类别：“疾病-疾病”、“疾病-部位”、“疾病-症状”、“治疗-疾病”、“治疗-症状”、“检查-疾病”和“检查-症状”，其详细定义如表2所示。

类别	定义
疾病	指导致病人身体或心理上出现的非正常现象，或者由医生对病人做出的诊断，并且是可以治疗的。
症状	泛指由疾病或其它突出状况导致身体不适或异常的表现，通常是病人主观感觉的不适或病理改变。
部位	指人的身体部位，包括器官或器官组成、身体系统以及身体位置或区域等。
检查	为了确认疾病是否存在，或是为了解疾病的病因而进行的检查项目、查体以及实施的检查设备。
治疗	针对疾病或症状而采取的药物、手术或医疗设备等治疗方法。

Table 1: 实体各类别定义

4.3 数据标注过程

为了获取高质量标注电子病历数据，在与医学专家讨论并制定中文电子病历实体及关系标注规范的基础上，选取了10,000份电子病历文档进行人工标注。标注过程可以分为两步：预标注和正式标注。预标注使用了20%的数据，每份电子病历数据采取两人同时标注的策略，标注结束后双方交互验证，对于标注相同的结果初步认定是准确的，反之交由医学专家进一步分析

关系类别	定义
疾病-疾病	泛指疾病之间的相关并发症、疾病表明疾病或者疾病的别名等。
疾病-部位	泛指疾病体现的部位，一般指发病部位，也有转移部位等情况。
疾病-症状	泛指疾病的一种体现形式，一般指疾病导致的某种症状。
治疗-疾病	泛指治疗应用于疾病，使疾病好转或恶化，或是没有提及治疗效果。
治疗-症状	泛指针对某些症状采取的治疗手段，或是因治疗所产生的症状。
检查-疾病	泛指检查确认了疾病的发生，或为证实疾病而采取某种检查手段。
检查-症状	泛指检查显示正常或者异常症状，或者检查确认是否存在症状。

Table 2: 关系各类别定义

评判，并找出标注结果不同的原因，以此进一步修订、细化电子病历标注规范。预标注结束后，正式标注严格按照制定的电子病历标注规范进行执行，同样地，每一份电子病历数据由两人同时标注，对于标注不同的结果，则分配给另外的标注者进行标注，直至标注结果相同。最终经过筛选去重后共获得实体30,058个，关系17,904对，各类别标注数量如表3、4所示。为了方便模型训练，标注完成后对较长的电子病历文本进行拆分，最终得到12,219条可训练样例。

4.4 数据集质量评估

为了保证数据集的质量，随机抽取了2,000个标注实体及2,500个标注关系进行质量评估，在评估阶段有医学专家全程参与，评估结果如表3、4所示，可以看出，实体和实体关系的准确率分别为92.9%和95.6%。另外，在评估过程中可以发现，实体的错误主要集中在实体的类型错误，即实体本身是正确的，但实体的类型是错误的，比如：“头痛”既可以充当疾病实体也可以充当症状实体，在不同的电子病历文本中可能表现出不同的实体类型；类似的，关系类别的错误也集中在实体类型的错误上，如“疾病-疾病”与“疾病-症状”。

实体类别	总数	抽样数	错误数	准确率(%)
疾病	12542	600	45	92.5
症状	7767	400	28	93.0
部位	843	100	4	96.0
检查	4646	400	31	92.3
治疗	4260	500	35	93.0
All	30058	2000	143	92.9

Table 3: 实体数量统计及质量评估

实体类别	总数	抽样数	错误数	准确率(%)
疾病-疾病	2935	200	8	96.0
疾病-症状	4221	200	13	93.5
疾病-部位	855	100	3	97.0
治疗-疾病	3388	500	15	97.0
治疗-症状	1938	500	27	94.6
检查-症状	1947	500	20	96.0
检查-疾病	2620	500	25	95.0
All	17904	2500	111	95.6

Table 4: 关系数量统计及质量评估

5 实验与结果分析

5.1 数据集与评价指标

基于中文电子病历的实体及关系抽取数据集CEMRIE总共包含12,219条样本，数据集按照80%:20%的比例划分为训练集和测试集，其详细数据如表5所示。采用严格匹配的方式来评估抽取三元组的效果，即当且仅当预测的三元组中，实体边界、实体类型以及关系类型完全正确时，预测的三元组才被视为正确；实验使用准确率P、召回率R以及F1值，并采用微平均(Micro-F1)的方式对结果进行评估。

CEMRIE	样本数	实体数	关系数
训练集	9773	24077	14349
测试集	2446	5981	3555

Table 5: CEMRIE数据集详情

5.2 超参数设置及实验环境

采用RoBERTa$_{\text{base}}$预训练模型(Liu et al, 2019)，隐藏层向量维度设为768。采用Adam优化器，初始学习率设为2e-5，并且当训练轮次到达20、50和70时，学习衰减为原来的一半，训练100轮。批处理大小设置为4，随机种子设置为99。

实验环境：操作系统Ubuntu20.04LTS，显卡NVIDIA GeForce RTX3090GPU，PyTorch版本为1.7.1，Python版本为3.7.11。

5.3 PIAN联合抽取实体关系性能测试

5.3.1 标准数据集实验结果

Dataset	Model	PLM	NER			RE		
			P	R	F1	P	R	F1
ACE05	Tab2Seq(Wang and Lu, 2020)	$\text{ALBERT}_{\text{large}}$	-	-	89.5	-	-	64.3
	PURE(Zheng et al, 2021)	$\text{ALBERT}_{\text{large}}$	-	-	89.7	-	-	65.5
	PFN(Yan et al, 2021)	$\text{ALBERT}_{\text{large}}$	-	-	89.5	-	-	66.8
	UNIRE▲(Wang et al, 2021)	$\text{ALBERT}_{\text{large}}$	**89.9**	**90.5**	**90.2**	**72.3**	60.7	66.0
	PIAN(Ours)	$\text{ALBERT}_{\text{large}}$	89.5	90.1	89.8	69.9	**66.5**	**68.1**
	PIAN(Ours)	$\text{BERT}_{\text{base}}$	88.6	88.5	88.5	67.7	61.6	64.5
	-w/o BFIM	$\text{BERT}_{\text{base}}$	87.6	88.3	87.9	70.1	57.5	63.2
	-w/o SAM	$\text{BERT}_{\text{base}}$	88.3	88.4	88.3	68.1	60.4	63.9
	-w/o BFIM&SAM	$\text{BERT}_{\text{base}}$	87.5	88.3	87.9	69.1	59.0	63.6
ACE04	Tab2Seq(Wang and Lu, 2020)	$\text{ALBERT}_{\text{large}}$	-	-	88.6	-	-	59.6
	PURE(Zheng et al, 2021)	$\text{ALBERT}_{\text{large}}$	-	-	88.8	-	-	60.2
	PFN(Yan et al, 2021)	$\text{ALBERT}_{\text{large}}$	-	-	89.3	-	-	62.5
	UNIRE▲(Wang et al, 2021)	$\text{ALBERT}_{\text{large}}$	88.9	**90.0**	89.5	67.3	59.3	63.0
	PIAN(Ours)	$\text{ALBERT}_{\text{large}}$	**89.6**	89.9	**89.8**	**70.2**	**58.8**	**64.0**
CMeIE	Baseline(Zhang et al, 2021)	$\text{RoBERTa}_{\text{large}}$	-	-	-	-	-	55.9
	CopyR_{RL}(Zeng et al, 2019)	$\text{BERT}_{\text{base}}$	-	-	-	54.0	55.7	54.6
	CasRel(Wei et al, 2020)	$\text{BERT}_{\text{base}}$	-	-	-	58.4	58.0	58.1
	NPCTS(王泽儒和柳先辉, 2022)	$\text{BERT}_{\text{base}}$	-	-	-	59.3	57.6	58.4
	PIAN(Ours)	$\text{RoBERTa}_{\text{base}}$	-	-	-	**63.8**	**58.8**	**61.2**

Table 6: PIAN在标准数据集上的实验结果，其中▲表示该方法利用了跨句信息；由于CMeIE为线上测评，无法提供实体结果

为了验证PIAN在实体关系联合抽取上的有效性，首先在标准数据集ACE04[0] (Doddington et al, 2004)，ACE05(Christopher et al, 2005)和中文医学数据集CMeIE[1](Zhang et al, 2021)上，与当前较为先进的模型进行比较，如表6所示。在数据集ACE05上，PIAN在关系结果上取了较大提升，比当前结果最好的共享参数模型PFN提高了1.3%；实体识别也取得了相当的结果，F1值仅比UNIRE低0.4%，这是因为UNIRE利用了额外的跨句子信息进行训练，其更有利于实体的预测。而本文方法没有使用跨句子信息，因此实体效果提升不明显。在数据集ACE04上，PIAN超越了当前效果最好的统一编码模型UNIRE，实体结果提升了0.3%，关系结果提升了1.0%。在中文医学数据集CMeIE上，PIAN结果显著高于NPCTS模型，NPCTS是对CasRel模型的改进，是一种基于指针级联标注策略的联合抽取模型，其中CasRel与CopyR_{RL}的实验结果为(王泽儒和柳先辉, 2022)复现的结果。本文模型在不借助医学预训练语言模型的情况下，达到了较好的结果，验证了模型在中文医学语料上的有效性。

5.3.2 消融实验

为了验证本文提出的BFIM以及注意力模块的有效性，基于$\text{BERT}_{\text{base}}$预训练模型在ACE05测试集上进行消融实验，结果如表6所示。当模型移除了交互模块后，由于两个任务无法进行有效的交互，实体和关系的效果均有下降。当移除注意力模块后，尽管两个任务的网络可以交互，但缺少学习自身任务特征的神经网络，与任务特征相关性较强的特征无法通过注意力网络进行强化，进而影响两个任务特征的质量。当两个模块全部移除后，模型性能下降最明显。综上，消融实验表明了本文提出的BFIM以及注意力模块能够显著提升模型效果，验证了PIAN模型的有效性。

5.4 PIAN在中文电子病历上的实验结果

5.4.1 CEMRIE数据集基准结果

将PIAN模型应用于中文电子病历上进行实体关系联合抽取，实体识别结果如表7所示，

[0]ACE04与ACE05语料来自于新闻文章、网上论坛等多种资源，共定义7种实体类型以及6种关系类型。
[1]CMeIE语料源于儿科及百种常见疾病训练语料，共定义11种医学实体类型以及44种关系类型。

实体类别	P(%)	R(%)	F1(%)
疾病	91.6	91.0	91.3
症状	81.5	82.7	82.1
部位	76.9	80.5	78.7
检查	94.5	93.1	93.8
治疗	86.9	84.3	85.6
All	**90.4**	**90.1**	**90.2**

Table 7: 实体各类别效果

实体类别	P(%)	R(%)	F1(%)
疾病-疾病	74.9	73.0	74.0
疾病-症状	77.5	77.0	77.3
疾病-部位	74.2	74.3	74.2
治疗-疾病	76.2	76.0	76.1
治疗-症状	70.3	72.4	71.3
检查-疾病	86.9	84.3	85.6
检查-症状	82.8	81.3	82.0
All	**80.8**	**80.6**	**80.7**

Table 8: 关系各类别效果

实体识别各个类别均取得了较好的效果，总体达到了90.2%的F1值。从结果可以看出，“症状”和“部位”类型的结果较差，其原因之一是“症状”类型实体与“疾病”类型实体在语义上较为相近，例如“偏头痛”属于疾病，但是容易与症状“头痛”混淆。“部位”类型的实体占总实体比例较少，模型无法充分学习其特征，若遇到复杂“部位”实体，如“输尿管跨越髂血管”、“肾盂输尿管连接部”等，则无法正确抽取。“疾病”、“治疗”和“检查”类型的实体由于特征较为明显且样本充足，抽取效果较好。

关系各类别的抽取结果如表8所示，可以看出关系抽取总体达到了80.7%的F1值。其中“疾病-症状”、“治疗-症状”以及“疾病-疾病”类别的效果较差，其原因是“症状”类型实体在语义上与“疾病”类型实体类似，模型不易区分，如果“疾病”(或“症状”)类型实体被错误预测为“症状”(或“疾病”)类型，关系抽取模型就会使用这些错误信息导致预测出错误的关系类型，所以与“症状”有关的关系类型效果较差。由于一部分“疾病”类型实体被预测为“症状”实体，因此会影响“疾病-疾病”类型关系效果。

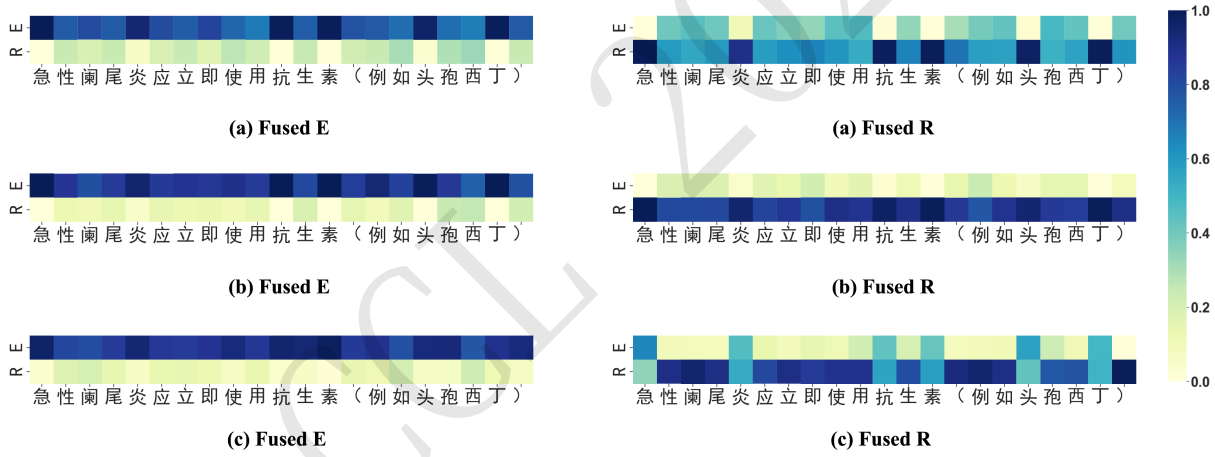

Figure 2: 双向特征交互模块中各个融合单元融合比例

5.4.2 双向特征交互模块分析

为了进一步探究BFIM的作用原理，将交互模块中的融合分数可视化，以更加直观地研究两个任务交互的具体信息。对于例句“急性阑尾炎应立即使用抗生素（例如头孢西丁）”，其中包含疾病实体“急性阑尾炎”、治疗实体“抗生素”和“头孢西丁”，其中疾病与两个药物实体存在“治疗-疾病”关系。

图2展示了模型中三个BFIM在生成融合特征时，两个任务特征融合的比例，其中Fused E/R表示融合后的特征，E/R表示分别来自实体任务和关系任务的特征比例。如图2.(a)所示，在第一个BFIM中，两个任务都捕获到了实体的头尾信息。如图2.(b)所示，第二个BFIM执行了较少的特征交互，这是因为在经过第一次交互后，每个任务都通过各自的注意力网络学习到了具有自身任务特点的特征，为了避免两种不同的特征混淆，该交互模块最大程度的保留了其原始的输入特征。如图2.(c)所示，在生成关系任务的融合特征时，来自实体识别任务的实体开始字符特征以较大比例融合到了对应的关系任务字符中，这表明实体的开始字符特征对关系

抽取任务具有重要作用；同时实体任务特征也融合了来自关系任务的特征。通过以上分析可知，BFIM能够从另一任务发现对自身任务有价值的信息，并通过自适应地学习一个最佳的融合特征比例实现充分的信息交互。

6 结论

本文提出了一种基于中文电子病历的平行交互注意力网络(PIAN)用于联合抽取实体及关系。为了充分利用实体识别和关系抽取两个任务的相关性，提出一个双向特征交互模块，其可以自适应地使一个任务中每个字符的特征与另一任务对应的特征动态融合，实现了双向细粒度的交互方式。模型在多个标准数据集上达到了最优结果，验证了本文方法的有效性。鉴于当前中文医学实体关系标注语料十分稀缺，与医学专家研讨并制定了标准的数据集标注规范，构建了中文电子病历实体关系抽取数据集CEMRIE并检验了数据集的质量，提供了基准结果。

参考文献

Bekoulis Giannis, Deleu Johannes, Demeester Thomas and Develder Chris. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34-45.

Christopoulou Fenia, Tran Thy Thy, Sahu Sunil Kumar, Miwa Makoto and Ananiadou Sophia. 2020. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39-46.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. The automatic content extraction (ace) program-tasks, data, and evaluation. *Journal of biomedical informatics*, 45(5):57-45.

Devlin Jacob, Chang Ming-Wei, Lee Kenton and Toutanova Kristina. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.

Doddington George R, Mitchell Alexis, Przybocki Mark A, Ramshaw Lance A, Strassel Stephanie M and Weischedel Ralph M. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. *Lrec*, 2(1):837-840.

Fei Hao, Zhang Yue, Ren Yafeng and Ji Donghong. 2021. A span-graph neural model for overlapping entity relation extraction in biomedical texts. *Bioinformatics*, 37(11):1581-1589.

Guo Menghao, Liu Zhengning, Mu Taijiang and Hu Shimin. 2021. Beyond self-attention: External attention using two linear layers for visual tasks. *arXiv preprin:2105.02358*.

He Kaiming, Zhang Xiangyu, Ren Shaoqing and Sun Jian 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

Lan Yinyu, He Shizhu, Liu Kang, Zeng Xiangrong, Liu Shengping and Zhao Jun. 2021. Path-based knowledge reasoning with textual semantic information for medical knowledge graph completion. *BMC Medical Informatics and Decision Making*, 21(9):1-12.

Lan Zhenzhong, Chen Mingda, Goodman Sebastian, Gimpel Kevin, Sharma Piyush and Soricut Radu. 2019. Albert: a lite bert for self-supervised learning of language representations. *International Conference on Learning Representations*.

Lee Jinhyuk, Yoon Wonjin, Kim Sungdong, Kim Donghyeon, Kim Sunkyu, So Chan Ho and Kang Jaewoo. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234-1240.

Li Fei, Zhang Meishan, Fu Guohong and Ji Donghong. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics*, 18(1):1-11.

Liu Shengyu, Tang Buzhou, Chen Qingcai and Wang Xiaolong. 2016. Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, 2016.

Liu Yinhan, Ott Myle, Goyal Naman, Du Jingfei, Joshi Mandar, Chen Danqi, Levy Omer, Lewis Mike, Zettlemoyer Luke and Stoyanov Veselin. 2019. Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo Ling, Yang Zhihao, Yang Pei, Zhang Yin, Wang Lei, Lin Hongfei and Wang Jian. 2018. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381-1388.

Luo Ling, Yang Zhihao, Cao Mingyu, Wang Lei, Zhang Yin and Lin Hongfei. 2020. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of biomedical informatics*, 103:103384.

Miwa Makoto and Bansal Mohit. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1105-1116.

Park Chanhee, Park Jinuk and Park Sanghyun. 2020. Agcn: attention-based graph convolutional networks for drug-drug interaction extraction. *Expert Systems with Applications*, 159:113538.

Rei Marek, Crichton Gamal KO and Pyysalo Sampo. 2016. Attending to characters in neural sequence labeling models. *arXiv preprint arXiv:1611.04361*.

Ren Feiliang, Zhang Longhui, Yin Shujuan, Zhao Xiaofeng, Liu Shilei, Li Bochao and Liu Yaduo 2021. A novel global feature-oriented relational triple extraction model based on table filling. *Proceedings of the Empirical Methods in Natural Language Processing*, 2646-2656.

Sui Dianbo, Chen Yubo, Zhao Jun, Jia Yantao, Xie Yuantao and Sun Weijian. 2020. Feded: federated learning via ensemble distillation for medical relation extraction. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2118-2128.

Sun Cong, Yang Zhihao, Su Leilei, Wang Lei, Zhang Yin, Lin Hongfei and Wang Jian. 2020. Chemical–protein interaction extraction via Gaussian probability distribution and external biomedical knowledge. *Bioinformatics*, 36(15):4323-4330.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Lukasz and Polosukhin Illia. 2017. Attention is all you need. *Proceedings of the Conference and Workshop on Neural Information Processing Systems*, 5998-6008.

Wang Jue and Lu Wei. 2020. Two are better than one: joint entity and relation extraction with table-sequence encoders. *Proceedings of the Empirical Methods in Natural Language Processing*, 1706-1721.

Wang Yucheng, Yu Bowen, Zhang Yueyang, Liu Tingwen, Zhu Hongsong and Sun Limin. 2020. Tplinker: single-stage joint extraction of entities and relations through token pair linking. *Proceedings of the International Conference on Computational Linguistics*, 1572-1582.

Wang Yijun, Sun Changzhi, Wu Yuanbin, Zhou Hao, Li Lei and Yan Junchi. 2021. Unire: a unified label space for entity relation extraction. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 220-231.

王泽儒, 柳先辉. 2022. 基于指针级联标注的中文实体关系联合抽取方法. 武汉大学学报(理学版), 1-7.

Wei Zhepei, Su Jianlin, Wang Yue, Tian Yuan and Chang Yi. 2020. A novel cascade binary tagging framework for relational triple extraction. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1476-1488.

Yan Zhiheng, Zhang Chong, Fu Jinlan, Zhang Qi and Wei Zhongyu. 2021. A partition filter network for joint entity and relation extraction. *Proceedings of the Empirical Methods in Natural Language Processing*, 185-197.

Yi Zibo, Li Shasha, Yu Jie, Tan Yusong, Wu Qingbo, Yuan Hong and Wang Ting. 2017. Drug-drug interaction extraction via recurrent neural network with multiple attention layers. *International Conference on Advanced Data Mining and Applications*, 554-566.

Zeng Daojian, Liu Kang, Lai Siwei, Zhou Guangyou and Zhao Jun. 2014. Relation classification via convolutional deep neural network. *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 2335-2344.

Zeng Xiangrong, He Shizhu, Zeng Daojian, Liu Kang, Liu Shengping and Zhao Jun. 2019. Learning the extraction order of multiple relational facts in a sentence with reinforcement learning. *Proceedings of the Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, 367-377.

Zhang Ningyu, Chen Mosha, Bi Zhen, Liang Xiaozhuan, Li Lei, Shang Xin, Yin Kangping, Tan Chuanqi, Xu Jian, Huang Fei and others. 2021. Cblue: a Chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Zheng Suncong, Wang Feng, Bao Hongyun, Hao Yuexing, Zhou Peng and Xu Bo. 2017. Joint extraction of entities and relations based on a novel tagging scheme. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1227-1236.

Zhong Zexuan and Chen Danqi. 2021. A frustratingly easy approach for entity and relation extraction. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 50-61.