基于多通道自注意力机制的电子病历实体关系抽取

宁尚明[1] 滕 飞[1),2)] 李天瑞[1),2)]

[1)](西南交通大学信息科学与技术学院 成都 611756)

[2)](西南交通大学人工智能研究院 成都 611756)

摘 要 电子病历是临床治疗过程中患者病情及治疗流程的重要载体之一,其中各类实体间关系包含了大量与患者健康相关的医学信息.因此,对电子病历文本的深度挖掘是获取医学知识、分析患者病情的有效手段之一.实体的高密度分布以及实体间关系的交叉互联为电子病历实体关系的抽取带来了极大挑战,应用于通识领域的实体关系抽取方法也因此受到极大的限制.针对这一文本差异性,本文提出一种基于多通道自注意力机制的"recurrent+transformer"神经网络架构,相比于主流的"recurrent+CNN"架构,该架构可强化模型对句级别语义特征的捕捉,提升对电子病历专有文本特点的学习能力,同时显著降低模型整体复杂度.此外,本文提出在该网络架构下的两种基于权重的辅助训练方法:带权学习的交叉熵损失函数以及基于权重的位置嵌入,前者用于缓解实体关系类别不均衡所造成的训练偏置问题,从而提升模型在真实分布数据中的普适性,同时可加速模型在参数空间的收敛速率;后者则用于进一步放大文本字符位置信息的重要性,以辅助提升 transformer 网络的训练效果.对比实验选用目前主流方法的 6 个模型作为基线,相继在 2010 i2b2/VA 及 SemEval 2013 DDI 医学语料中进行验证.相较于传统自注意力机制,多通道自注意力机制的引入在模型整体 $F1$ 指标中最高实现 10.67%的性能提升,在细粒度单项对比实验中,引入类别权重的损失函数在小类别样本中的 $F1$ 值最高提升近 23.55%.

关键词 关系抽取;电子病历;多通道自注意力;recurrent+transformer;语义特征

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2020.00916

Multi-Channel Self-Attention Mechanism for Relation Extraction in Clinical Records

NING Shang-Ming[1] TENG Fei[1),2)] LI Tian-Rui[1),2)]

[1)](*School of Information Science and Technology, Southwest Jiaotong University, Chengdu* 611756)

[2)](*Institute of Artificial Intelligence, Southwest Jiaotong University, Chengdu* 611756)

Abstract The electronic medical record is one of the important carrier for patient's condition and treatment during the clinical treatment process. The relationship between various types of entities contains a large number of medical knowledge related to the information of the patient. Therefore, the deep mining of electronic medical records is one of the effective means to obtain medical knowledge and analyze the patient's condition. The high-density distribution of entities and the cross-connection of relationships between entities pose great challenges for the relation extraction in electronic medical records. For that, the methods of relation extraction applied in the general fields are greatly limited. In view of the characteristics of that, this paper proposes a "recurrent+transformer" architecture with multi-channel self-attention mechanism to enrich the semantic features of the sentence level, thus improving the learning ability of the characteristics for electronic medical records and reducing model complexity. In addition, this paper also proposes two auxiliary training methods based on weight, which are weighted-based cross entropy loss function

收稿日期:2019-09-01;在线出版日期:2020-02-08.本课题得到国家自然科学基金(61572407)和四川省科技计划(2017SZYZF0002)资助.
宁尚明,硕士研究生,主要研究方向为自然语言处理.E-mail:ningsm624@163.com.**滕 飞**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究领域为云计算、数据挖掘.E-mail:fteng@swjtu.edu.cn.**李天瑞**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为人工智能、数据挖掘、粗糙集与粒计算、云计算与大数据.

and weighted-based position embedding. The former is applied to avoid the problem of training bias caused by categories imbalance, thus improving the universality of the model in the real distribution and accelerating the convergence rate. The later enhances the importance of position information with each character, which helps improve the training effect of transformer network. We selected six models with the best performance in the two methods as the baselines, and verified them in the 2010 i2b2/VA and SemEval 2013 DDI medical corpus. Compared with the traditional self-attention mechanism, the highest performance improvement of 10.67% is achieved in the overall $F1$ score of the model with the multi-channel self-attention mechanism. In the fine-grained single-item comparison experiment, the weighted-based loss function increases the $F1$ value in the small category sample by nearly 23.55%.

Keywords　relation extraction; electronic medical record; multi-channel self-attention; recurrent + transformer; semantic features

1　引　言

电子病历是指医务人员在医疗活动过程中，使用医疗机构信息系统生成的文字、符号、图表、图形、数据、影像等数字化信息，并能实现存储、管理、传输和重现的医疗记录①. 电子病历中诸如疾病、症状、治疗等实体是整个治疗流程中的核心信息，而各独立实体间的关系则是医疗知识的直接表达. 因此，通过对电子病历文本的分析与挖掘是医疗知识获取的重要手段之一，该认知已得到广泛认可与实施[1]. 结合自然语言处理相关技术，围绕电子病历等医学文本的相关研究，可为智能导诊、患者问答等场景提供有效的支持与应用.

实体关系抽取作为自然语言处理底层任务的分支之一，旨在从非结构化文本中识别各实体间的关系，进而为语料库构建、知识图谱构建等高层任务提供支持[2]. 目前，面向通识领域的关系抽取研究进展较为迅速，然而，受限于医学领域知识及开放数据集规模，以电子病历文本为核心的实体关系抽取研究仍面临极大困难与挑战.

面向电子病历的实体关系抽取方法，其技术难点在于高密度的实体分布以及交叉互联的实体关系，该特性主要表现于单句中包含多个实体，实体可能隶属于不同类别，实体间产生多种类型关系，并且同一实体因交叉会参与生成多个关系对[3-4]. 以2010 i2b2/VA 语料为例："The patient had bone marrow biopsy, for underline persistent pancytopenia, revealing mild hypercelluarity with 80% leukemic cells."（该患者进行了骨髓活检，发现持续性全血细胞减少，显示轻度细胞增生，伴有80%白血病细胞）该句包含两种类型的 4 个实体，实体间共产生 4 组关系，分别为[骨髓活检，持续性全血细胞减少]=TeCP，[骨髓活检，轻度细胞增生]=TeCP，[骨髓活检，80%白血病细胞]=TeCP，[轻度细胞增生，80%白血病细胞]=PIP（具体关系释义请参照表 1）. 上述例子可以明显看出，简短的病程记录包含高密度的实体分布，并且同一实体[骨髓活检]相继参与到三组关系中. 此外，本文对 2010 i2b2/VA 语料进行统计后发现，平均每 18 个字符（包含实体字符）将出现 4 个不同的实体，且平均发生 6 组实体间关系. 而通识领域实体关系抽取公开语料，如 SemEval2010_task8 中，平均每 23 个字符（包括实体字符）包含 2 个不同实体，且平均仅产生 1 组关系. 该统计结果表明，言简意赅的电子病历短文本中有较高密度的实体分布，且实体间交错产生多种关系. 句中高密度的实体分布看似能够为模型拟合提供更为丰富的信息，但对于实体关系抽取任务来说，同一实体多次参与不同关系对的组成，且各关系可能隶属不同类别. 因此，在仅有的标注信息支撑下，一旦模型缺乏句级别语义信息的表征能力，将易导致对此类交错关系的欠拟合，最终影响电子病历关系抽取的性能表现. 因此，上述文本特点要求模型更善于捕捉与理解语义层面的特征，而并非单一的时态或短语特征.

早期的研究成果大多基于传统统计学习方法[5-6]，这些方法的共性缺点是前期需要进行大量的特征工程以抽取有效的特征集，如词性、最短依赖路径甚至是设计具体的核函数[7]. 换言之，人工特征抽

① 中华人民共和国卫生部. 电子病历基本规范（试行）. http://www.gov.cn/zwgk/2010-03/04/content_547432.htm, 2013, 12, 7.

取的质量将决定模型性能的上限.近年来,神经网络及深度学习相关技术被逐渐应用在关系抽取任务中,并在通识领域取得较大的突破.该类方法的最大优势在于对特征工程的裁剪,在提升建模效率的同时也带来较大幅度的性能提升[8-9].以循环神经网络和卷积神经网络为代表的深度学习架构目前已在电子病历实体关系抽取任务中得到成功应用[10-13].在最新的研究成果中,注意力机制也被成功迁移至此类任务中,并在模型训练效率及特征抽取方面取得进一步突破[14].然而,这些成果仅是通识领域模型向电子病历文本的简单迁移,尚未真正解决电子病历中高密度的实体分布以及复杂的实体关系对所造成的信息干扰.

根据上述对电子病历文本特点的分析及最新研究方法的比对,本文摒弃该任务最佳基线模型的"recurrent+CNN"网络架构,提出一种复杂度更低,分类效果更为显著的"recurrent+transformer"模型架构.其 transformer 组件引入多通道自注意力机制,以实现对电子病历文本句级别语义特征的深入挖掘与学习;recurrent 层沿用 BiLSTM 网络,用于对电子病历文本浅层语义特征的捕捉与学习.本文的主要贡献如下:

(1)一种更为高效的神经网络架构."recurrent+CNN"是当前医学文本领域实体关系抽取任务的主流建模方法,为进一步建模表征更全面的句级别语义特征,缓解电子病历实体关系交错关联等难题,并考虑整体网络的训练效率,本文提出复杂度更低的"recurrent+transformer"网络架构,其 recurrent 层指代循环神经网络或其相关变种网络,transformer 层则由注意力机制具体实现.本文在大量对比实验中验证了该架构的有效性;

(2)一种更有效的自注意力机制:多通道自注意力机制.区别于传统自注意力机制,该方法通过学习多组权重向量来拟合更为丰富的句级别语义信息,从而提升模型对电子病历高密度实体分布以及复杂实体关系的特征学习能力.实验比对及注意力权重可视化结果表明,多通道注意力机制的引入有助于模型对句级别语义信息的捕捉与编码.

(3)两种基于权重的辅助训练方法.为进一步强化上述网络架构对电子病历实体关系抽取的建模效果,本文针对医学文本固有的类别不均衡所造成的学习偏置问题,在电子病历实体关系抽取任务中提出一种带权学习的交叉熵损失函数,该方法不仅有助于模型对小类别样本的拟合,同时可加速模型收敛速率.此外,为进一步提升位置信息对 transformer 结构的训练影响,本文首次提出一种面向电子病历文本的基于权重的位置嵌入方法,在充分利用文本字符位置信息的同时,放大目标实体附近文本的重要性,并削减远距离文本对模型训练的影响.

2 相关工作

实体关系抽取的相关研究成果经历了从统计学习方法到深度学习模型的演进.在早期的相关研究成果中,以特征工程为核心的机器学习模型成为实施该任务的主流方法[2,6].其中,以构建有效核函数为建模思路的支持向量机模型是较为通用的方法之一[15-17].此类方法所使用的基模型虽然有较为完备的理论支撑,但却依赖大量人工干预,如通过繁琐的特征工程来筛选最为有效的特征集供模型学习.此外,为支持向量机设计有效的核函数同样是一项耗时费力的工程.

以神经网络为支撑的深度学习方法为通识领域关系抽取任务提供了一种新的解决思路,无需进行大量的特征构造与筛选,甚至仅依赖原始文本便可达到与机器学习方法同样的效果[18-19].其中最为典型的代表为循环神经网络(recurrent)及卷积神经网络(CNN)的系列架构[20-22],这些方法仅使用基础神经网络模型将关系抽取任务转化为分类问题进行建模.RNN 的使用有助于对文本序列前后依赖信息的学习,但缺少对句法及语义层面的特征挖掘.加入 CNN 框架的建模方法旨在关注文本序列的局部特征,但同时损失了对全局信息的把控.为了同时考虑上述单一模型存在的缺陷,BRCNN 将二者进行结合,使用双向循环神经网络捕捉文本序列前后信息,进而引入 CNN 卷积操作进一步捕捉文本局部特征[23].尽管 BRCNN 能够同时考虑双向文本信息以及局部特征,但其仍受限于单一模型的性能而缺乏对文本语义层面的深度挖掘.Peng 等抛弃 CNN 结构,采用双向循环神经网络与注意力机制相结合的建模方法来进一步对句级别特征进行学习与编码[24].

近年来,以深度学习方法为支撑的相关理论同样为电子病历实体关系抽取问题提供了更有效的解决思路.Sahu 等以 CNN 为基模型对该问题进行建模并在电子病历文本上进行尝试[10],作者认为,CNN 所提取的局部特征有助于表达高密度的实体分布特性,但模型对大量相距较远的实体对将缺乏判断力度.He 等将 CNN 与最大池化层(max poo-

ling)结合作为核心网络架构，并针对电子病历实体关系分布不均衡的特点，推导出基于类别约束矩阵的惩罚项，与损失函数一并对模型进行学习和训练[25]，该方法仍可视为卷积神经网络的简单应用，并仅在 I2B2 数据集中进行验证，因此其模型的普适性还有待探讨. BLSTM 是循环神经网络在该任务上的成功应用之一[12]，作者使用双向循环神经网络 BiLSTM 对文本信息进行特征抽取与建模，然后分别通过最大池化层与注意力机制对隐层输出做进一步编码，进而通过向量拼接的方式传入 softmax 层进行模型学习与分类. 虽然该方法率先引入双向循环网络对该任务进行建模，但受限于 BiLSTM 对文本信息重要性的区分能力，该模型依旧无法适用于风格差异性明显的电子病历中. 此后，大量研究人员采用并验证了“recurrent＋CNN”结构在电子病历关系抽取任务中的有效性，该网络构建方法也同时成为主流基准模型. 例如，CRNN 将 CNN 与 RNN 相结合以兼顾文本局部及全局信息[12]，作者分别于 RNN 层及 CNN 层之后施加池化层来对冗余信息进行过滤，从而更专注于对文本短语特征的学习. 此外，文中同样对注意力机制进行了验证，即将 CNN 后的最大池化层替换为注意力层. 然而，受限于 RNN 梯度消失等缺陷，CRNN 对较长文本仍旧无法很好地进行依赖信息的学习.

事实上，随着以注意力机制为核心思想的 transformer 结构的提出与完善，一定程度上兼顾了长文本信息的学习能力，同时提升了模型整体的训练效率[26]. 然而，对于电子病历这样的领域专有文本，传统的自注意力机制依旧在句级别语义信息的捕捉上存在缺陷. 因此，本文针对电子病历文本的差异性及特点，同时考虑模型整体的训练复杂度，引入多通道自注意力机制，提出一种“recurrent＋transformer”的网络架构，用于提升实体关系抽取在电子病历文本上的性能表现.

3　模型介绍

本节介绍“recurrent＋transformer”架构的网络层组织结构，并着重描述以多通道注意力机制(multi-channel：MCatt)及 BiLSTM 网络为核心的 BLSTM-MCatt 模型结构及其工作原理. 图 1 详细展示了该模型整体结构及内部数据变换的流程，其中灰色实心框为已有技术的引用，波点填充框为本文所提出内容的具体实现. 共包含以下 5 个核心组件：

(1) 输入层(Input layer)：原始电子病历文本按空格切分作为输入.

(2) 多嵌入层(Multi-embeddings layer)：包含传统词嵌入层(word embedding)与本文所提出的基于权重的位置嵌入层(weighted position embedding)：词嵌入层的输入为预训练词向量，位置嵌入层的参数矩阵通过随机初始化参与模型训练，两种嵌入层的输出向量进行拼接作为原始文本低层特征的向量表示.

(3) 底层特征抽取器(Underlying-Level feature

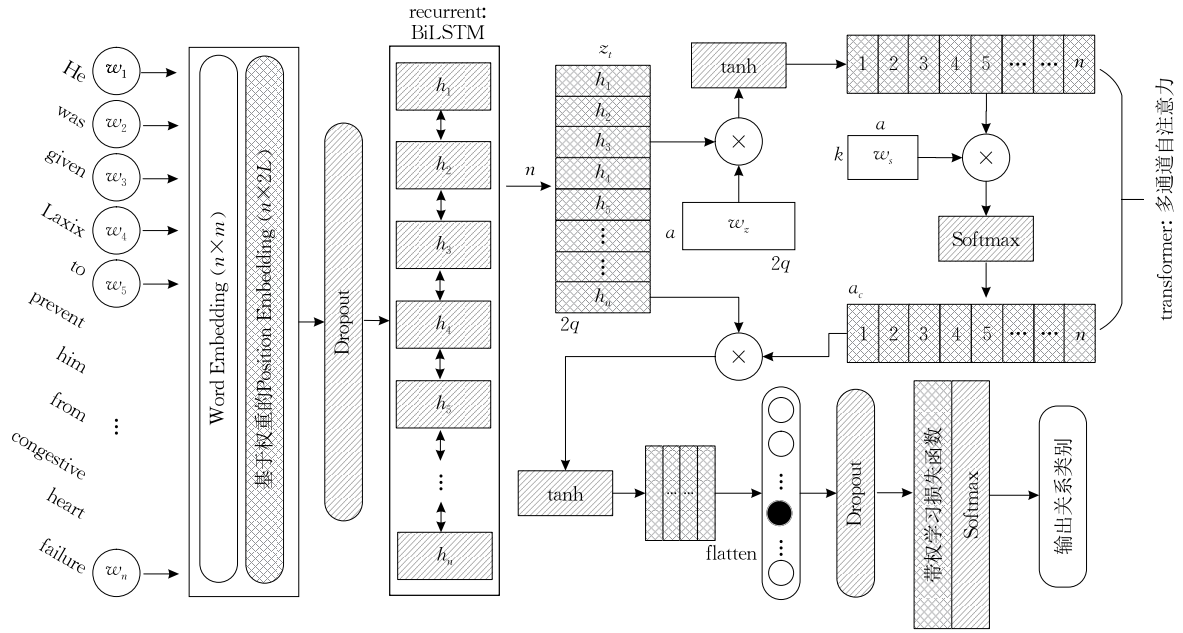

图 1　含有 k 通道的 BLSTM-MCatt 网络结构

extractor):BiLSTM 作为本文模型 recurrent 层的具体实现,作用于多嵌入层的输出,以捕捉文本序列的长短依赖特征.

(4)高层特征抽取器(High-Level feature extractor):引入多通道自注意力机制作为 BLSTM-MCatt 中 transformer 组件的具体实现,作用于 BiLSTM 网络之后,通过拟合多组权重向量来捕捉句级别高层特征.实验验证及注意力权重可视化结果表明,多通道自注意力机制在兼顾时态、短语等简单句法特征基础上,能够进一步提升文本各成分间重要性的区分度,从而有助于电子病历中复杂交错关系的分类与抽取.

(5)带权学习的损失函数(Weighted-based Loss Function):推导并制定出一套有效的类别权重计算方法,各类别权重作为参数向量与交叉熵损失函数共同参与训练.类别权重信息的引入,可摒弃人工采样所引入的随机误差,保证原生医学实体关系的固有分布.实验结果表明,该方法在改善医学实体关系类别分布不均衡问题的同时,加速了训练阶段模型整体的收敛速率.

3.1 Recurrent 层——BiLSTM

LSTM(Long Short-Term Memory)及其相关变种常被用于对序列数据的建模.由于文本数据可看作是具有前后依赖关系的序列数据,因此同样适用于 LSTM.在本文所提出的 BLSTM-MCatt 结构中,使用双向循环神经网络(BiLSTM)作为底层特征抽取器,对 multi-embedding 层的输出做粗粒度的特征抽取与编码.

LSTM 特殊的门控结构使其具有长短距离信息记忆的功能,其核心组件包含输入门 i_t、遗忘门 f_t 以及输出门 o_t,这些门控单元与记忆细胞(memory cell)c_t 相互配合,共同决定对隐藏层 h_t 信息的更新或是丢弃[27].BiLSTM 是 LSTM 的变体之一,其特点在于双层双向的前向计算与反向传播,式(1)即为 BiLSTM 在 t-th时刻的输出 z_t,其中使用对应位置求和的方式对双向隐层向量进行融合.

$$z_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}] \tag{1}$$

3.2 Transformer 层——多通道自注意力机制

以自注意力机制为核心的神经网络架构近年来逐渐成为自然语言处理领域的研究热点.传统自注意力机制通过学习一组权重向量来表达句中各字符的重要性,以此来捕捉句中的不同成份以及句法特征.然而该方法仅能对句中的简单短语、时态等明显成份进行学习,依旧缺乏对句级别语义特征的捕捉能力.对于实体分布密集,实体间关系交错出现的电子病历文本,需更多关注句中不同语义成份间的信息挖掘,因此本节引入多通道自注意力机制(MCatt),对句级别多语义信息进行表征[28].

"通道"是指对句子施加一次注意力机制并输出一维注意力权重向量,因此"多通道"是指 MCatt 将同时对目标句进行多次注意力操作,并产生多组权重向量.因此,区别于传统自注意力机制,MCatt 将输出一个 2 维权重矩阵用于表征句子的不同成分.基于该注意力权重计算方法,模型将学习并对句级别高层语义特征进行表征,有助于模型整体训练过程中对参数空间的快速搜索.对于高密度实体分布以及实体关系交叉关联的文本特点,上述方法所得到的句级别嵌入表征能够在增强语义特征表达的基础上,削弱冗余信息的重要性.因此,针对电子病历专有的文本特点,本文首次引入多通道注意力机制作为 transformer 层的具体实现,以提升实体关系抽取在电子病历中的建模效果.

多通道注意力机制的工作流程及原理如下:

(1)BiLSTM 将学习到的长短距离依赖信息进行拼接,并输出隐藏层向量 $\boldsymbol{Z}_t = \langle z_1, z_2, \cdots, z_n \rangle$,若设置单层 LSTM 隐层神经元个数为 q,则 $\boldsymbol{Z}_t$ 维度为 $n \times 2q$.

(2)BiLSTM 的输出 $\boldsymbol{Z}_t$ 将作为多通道自注意力的输入.对于传统自注意力机制,其注意力权重 α 可由式(2)与(3)计算得到,其中 $\boldsymbol{W}_z$ 和 $\boldsymbol{\omega}_s$ 分别是维度为 $a \times 2q$ 和 a 的可学习参数矩阵,a 在模型实现过程中可视为感知机参数,由用户设定.

$$M = \tanh(\boldsymbol{W}_z \cdot \boldsymbol{Z}_t^{\mathrm{T}}) \tag{2}$$

$$\alpha = \mathrm{softmax}(\boldsymbol{\omega}_s \cdot M) \tag{3}$$

根据上述传统自注意力机制的实施原理,MCatt 引入多通道概念,用于扩展模型对句中各成份信息的学习能力.给定通道数 C,MCatt 将在各通道 C_i 中分别进行一次自注意力权重计算,可看做是目标句子中第 i 种成份的重要性.因此,若句中关键信息较少,则通道数的增加会导致模型在训练过程中的冗余信息过多,从而影响下游任务的效果,本文也在实验章节验证了通道数对电子病历关系抽取的影响,同时也通过权重可视化的方式,直观感受多通道注意力机制对句中关键信息的捕捉能力.该过程可形式化为式(4),其中 $\boldsymbol{\omega}_s^{C_i}$ 指代通道 C_i 中的一组可学习权重矩阵,通道数 C 作为关键超参数,需根据实际业务场景进行相应调整.

$$\boldsymbol{\alpha}_{Ci} = \mathrm{softmax}(\boldsymbol{\omega}_s^{C_i} \cdot \tanh(\boldsymbol{W}_z \cdot \boldsymbol{Z}_t^{\mathrm{T}})) \tag{4}$$

各通道自注意力的计算相互独立，可通过矩阵运算实现多通道并行化. 若给定通道数 $C\in[1,k]$，式(4)中的参数 ω_s 可扩展为 $k\times a$ 维规模的 2 维矩阵 $\boldsymbol{W}_s$. 因此，可将 ω_s 替换为 $\boldsymbol{W}_s$ 并得到多通道自注意力权重的计算方式，如式(5)所示.

$$\boldsymbol{\alpha}_C = \mathrm{softmax}(\boldsymbol{W}_s \cdot \tanh(\boldsymbol{W}_z \cdot \boldsymbol{Z}_t^{\mathrm{T}})) \tag{5}$$

(3) 传统多头自注意力机制(multi-head attention)通过在 6 个相同的层级结构中随机初始化权重矩阵来进行多头学习，从而一定程度上避免大量冗余信息被模型学习. 在本文工作中，MCatt 通过加入约束项(6)来使得各通道的学习过程具有较好的区分度，其中 $\|*\|$ 代表矩阵的 Frobenius 范数，这样的计算方式可确保相同维度数字差异性越小，则惩罚力度越大，反之则对损失函数的惩罚力度变小，该惩罚项将作为损失函数的一部分对模型一同进行训练[27]. 因此，每增加一个通道，MCatt 将会对句中某一成分 i 进行权重拟合，即可看作各通道的学习结果分别代表句中的不同成份，本文 4.4 节的权重热力图展示了多通道自注意力机制对句中各成分的学习与区别能力. 对比 multi-head attention 的多层串行结构，带有约束项的多通道自注意力机制在过滤冗余信息的基础上，极大简化了注意力层的网络结构，一定程度上提升模型的训练效率.

$$Penalty = \|\boldsymbol{\alpha}_C \cdot \boldsymbol{\alpha}_C^{\mathrm{T}} - \boldsymbol{I}\| \tag{6}$$

(4) 在模型的具体实现过程中，本文使用两层感知机(perceptron)来计算式(5)的权重矩阵 $\boldsymbol{\alpha}_C$，最后通过式(7)与(8)，将 $\boldsymbol{\alpha}_C$ 与 $\boldsymbol{Z}_t$ 相乘再进行规范化得到自注意力层的输出 $\boldsymbol{H}^*$，即句级别语义特征的向量表征.

$$\boldsymbol{Z}_t^{att} = \boldsymbol{\alpha}_C \cdot \boldsymbol{Z}_t \tag{7}$$

$$\boldsymbol{H}^* = \tanh(\boldsymbol{Z}_t^{att}) \tag{8}$$

3.3　基于权重的辅助训练提升方法

本节以“recurrent＋transformer”为基础网络架构，使用 BiLSTM 与多通道注意力机制相结合的具体实现作为本文电子病历关系抽取任务的解决方案，在此基础上，为进一步考虑电子病历文本的差异与特征，并全面提升“recurrent＋transformer”架构在电子病历文本中的学习能力，提出两种基于权重的辅助训练提升方法：带权学习的损失函数以及基于权重的位置嵌入计算方法.

(1) 带权学习的损失函数

Softmax 分类器常被用作对隐藏层向量的概率映射，在本文模型中，同样沿用 softmax 层来判别句中目标实体 en_1 与 en_2 之间的关系 R_i. MCatt 输出 $\boldsymbol{H}^*$ 作为分类器的输入，由式(9)与(10)进行类别概率的计算与判别，其中 $\boldsymbol{W}^{(S)}$ 是 softmax 分类器的参数矩阵，$\boldsymbol{b}^{(S)}$ 为偏置参数矩阵.

Softmax 分类器常被用作对隐藏层向量的概率映射，在本文模型中，同样沿用 softmax 层来判别句中目标实体 en_1 与 en_2 之间的关系 R_i. MCatt 输出 $\boldsymbol{H}^*$ 作为分类器的输入，由式(9)与(10)进行类别概率的计算与判别，其中 $\boldsymbol{W}^{(S)}$ 是 softmax 分类器的参数矩阵，$\boldsymbol{b}^{(S)}$ 为偏置参数矩阵.

$$\hat{p}(y\,|\,[en_1,en_2]) = \mathrm{softmax}(\boldsymbol{W}^{(S)}\boldsymbol{H}^* + \boldsymbol{b}^{(S)}) \tag{9}$$

$$\hat{y} = \underset{y}{\mathrm{argmax}}\,\hat{p}(y\,|\,[en_1,en_2]) \tag{10}$$

通识领域中实体关系抽取任务可通过引入抽样算法来缓解类别不均衡所导致的训练偏置问题，然而在医疗领域，数据固有的正确分布往往是失衡的，例如常见病的发病率高于罕见病已是不争的事实，无论是科室规模或是病案数量，都造成电子病历数据分布不均衡的情形. 因此，在该领域数据中利用采样算法来平衡数据分布是违背自然规律的做法，在其上训练得到的模型不具备很好的鲁棒性. 基于上述分析，本节首次在该任务中提出基于类别权重的损失函数，以缓解类别分布不均衡问题.

Softmax 分类器所采用的常见损失函数为交叉熵(cross entropy loss)：

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log(\hat{y}_i) + \lambda\|\theta\|_F^2 \tag{11}$$

其中 $t\in\mathbb{R}^c$，代表当前样本所属类别的 onehot 编码，c 是类别总数，$\hat{y}_i$ 表示 softmax 层将隐层向量映射为各类别的概率值. 式(11)中最后一项代表 L2 正则项.

交叉熵损失函数的目的是为了计算训练样本当前的总体损失从而推动模型做进一步参数更新，本节提出通过在损失函数中引入类别权重 $\boldsymbol{W}^{wgt} = \langle w^{c_1}, w^{c_2}, \cdots, w^{c_c}\rangle$ 来调整各类别对总体损失的贡献比例，从而平衡小类别样本在训练过程中的重要性. 其类别权重的设计思路依赖各类别样本原始分布的状况，若类别 i 对应的样本数 N_i 小于全体类别样本数均值，则考虑为其赋予大于 1 的权重，进而在训练过程中使小类别样本误判惩罚得到放大，反之则赋予小于 1 的权值，其类别权重 w^{c_i} 的具体计算方式如式(12)所示.

$$w^{c_i} = \begin{cases} 1 + \dfrac{N_i^{\mathrm{reverse}}}{N^{\mathrm{all}}}, & N_i < \mathrm{avg}(N^{\mathrm{all}}); \\ 1 - \dfrac{N_i^{\mathrm{reverse}}}{N^{\mathrm{all}}}, & N_i \geq \mathrm{avg}(N^{\mathrm{all}}) \end{cases} \tag{12}$$

首先定义样本总数为 $N_{\mathrm{all}} = \sum_{i=1}^{c} N_i$，其中 $N_i(i\in$

$\mathbb{R}^c$)代表训练集中第 i 类样本总数,并按升序排序得到有序数组 D,则 N_i^{reverse} 代表 D 中下标为($c-i$)所对应的值,对 D 求均值得到 avg(N^{all}),将其直接代入式(12)方可得到各类别的相应权重 w^{c_i}. 最后,更新式(11)中的 t_i 为 $t_i \boldsymbol{W}^{wgt}$,并加入惩罚项(6),则得到式(13)即为本节所提出的带权学习的损失函数.

$$J_{wgt}(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \boldsymbol{W}^{wgt} \log(\hat{y}_i) + \lambda \|\theta\|_F^2 + \|\boldsymbol{\alpha}_C \cdot \boldsymbol{\alpha}_C^{\mathrm{T}} - \boldsymbol{I}\| \tag{13}$$

(2)基于权重的位置嵌入层

词嵌入方法是自然语言处理常用的字词表征手段,而位置嵌入方法(positionembedding)的优势和效果已在 transformer 架构的相关研究中被证实[24]. 因此,本文模型的多嵌入层由传统的词嵌入以及本节提出的基于权重的位置嵌入层所构成,其中带有权重信息的位置嵌入层将进一步强化实体附近字符的重要性,并削弱远距离字符的影响.

原始电子病历文本可看作由不同句子构成,句中 $\boldsymbol{S}=\langle w_1, w_2, \cdots, w_n\rangle$ 的每个字符 w_i 可通过嵌入表(embedding table) $\boldsymbol{W}^{\text{table}} \in \mathbb{R}^{m \mid V \mid}$ 被映射为一组唯一的离散数值向量 $\boldsymbol{e}^{w_i}$,其中 V 和 m 分别代表词表规模以及用户指定的词向量维度,那么 $\boldsymbol{e}^{w_i}$ 便可由式(14)计算得到,其中 $\boldsymbol{v}_i^{\text{onehot}}$ 是大小为 V 的 onehot 向量. 对于位置嵌入,首先直接算得各字符与目标实体之间的相对距离 $p_i^{en_j} \in [-L, L]$,其中 i 代表句中字符的相对位置,en_j 代表第 j 个实体,L 作为一项超参数代表用户所指定的相对距离的上限. 对句中所有字符进行相对距离计算后可得到两组位置向量:$\boldsymbol{S}_{en_1} = \langle p_1^{en_1}, p_2^{en_1}, \cdots, p_n^{en_1}\rangle$,$\boldsymbol{S}_{en_2} = \langle p_1^{en_2}, p_2^{en_2}, \cdots, p_n^{en_2}\rangle$,分别代表各字符到句中两个目标实体的相对距离所构成的向量. 类似于 word embedding,每一个相对距离 $p_i^{en_1}$ 或 $p_i^{en_2}$ 可通过式(15)被映射为唯一的离散值向量. 其中 $\boldsymbol{P}^{\text{table}} \in \mathbb{R}^{n \mid 2L \mid}$ 指代位置嵌入表,n 代表用户指定的位置向量维度,类似的,$\boldsymbol{r}_i^{\text{onehot}}$ 是大小为 $2L$ 的 onehot 向量.

$$\boldsymbol{e}^{w_i} = \boldsymbol{W}^{\text{table}} \cdot \boldsymbol{v}_i^{\text{onehot}} \tag{14}$$

$$\boldsymbol{e}_i^{en_j} = \boldsymbol{P}^{\text{table}} \cdot \boldsymbol{r}_i^{\text{onehot}} \tag{15}$$

本节在传统位置嵌入方法的基础上,提出基于权重的位置嵌入,即为各字符所对应的 $\boldsymbol{e}_i^{en_j}$ 赋予相应的权重 $|L/p_i^{en_j}|$. 因此,式(15)可更新为式(16). 该权重的引入将对位置信息起到放缩作用,针对实体密度较大的电子病历文本来说,该权重信息的引入,一定程度上剔除了相对距离较远字符带来的冗余信息,同时提升了实体附近字符的重要性.

$$\boldsymbol{e}_{i_{wgt}}^{en_j} = \frac{\boldsymbol{P}^{\text{table}} \cdot \boldsymbol{r}_i^{\text{onehot}}}{|\ p_i^{en_j}/L\ |} \tag{16}$$

最终,多嵌入层将词嵌入与权重位置嵌入层的向量进行拼接作为最终输出,即式(17)所示.

$$emb_s = \langle \boldsymbol{e}^{w_1} \cdots \boldsymbol{e}^{w_n} \mid \boldsymbol{e}_{1_{wgt}}^{en_1} \cdots \boldsymbol{e}_{n_{wgt}}^{en_1} \mid \boldsymbol{e}_{1_{wgt}}^{en_2} \cdots \boldsymbol{e}_{n_{wgt}}^{en_2} \rangle \tag{17}$$

4 实验结果与分析

本节将分别对三处主要贡献(多通道自注意力机制、带权学习的损失函数以及基于权重的位置嵌入)进行验证与比对分析. 实验结果表明,本文所提出的模型在电子病历公开数据集上的验证效果超越了已有基线模型.

4.1 数据集与实验设置

分别选用 2010 i2b2/VA 关系语料库[29]与 SemEval 2013 Task-9 DDI[30-32]数据集进行模型验证. 前者是公认的电子病历实体关系评测数据集之一,表 1 详细描述了其关系类别含义及相关统计信息. 后者 DDI 数据集包含 Medline 以及 DrugBank 数据库,表 2 展示了 DDI 中所包含实体关系的含义及相关统计信息.

表 1 I2B2 语料集相关信息

关系	定义	训练集	测试集
TrIP	治疗改善医疗问题	165	41
TrWP	治疗恶化医疗问题	109	26
TrCP	治疗导致医疗问题	436	108
TrAP	针对医疗问题进行治疗	2131	532
TrNAP	因医疗问题而不进行治疗	140	34
TeRP	检验显示医疗问题	2457	614
TeCP	进行检验以查证医疗问题	409	101
PIP	医疗问题表明医疗问题	1776	443
None	无关系	52211	13045
总和	/	59834	14944

表 2 DDI 数据集相关信息

关系	定义	训练集	测试集
Mechanism	药物代谢动力学机制	1264	302
Effect	药效的相互作用	1620	360
Advice	同时用两种药的相关意见	820	221
Int	无任何信息的药物交互	140	96
None	无关系	12651	3046
总和	/	16495	4025

2010 I2B2/VA:该数据集源于三所医院的出院小结(dischargesummaries),其中共包含八种实体关系:治疗引起的医疗问题(TrCP)、治疗手段管控医疗问题(TrAP)、治疗恶化医疗问题(TrWP)、治疗改善/治愈医疗问题(TrIP)、由于医疗问题而未

给予治疗(TrNAP)、检验表明医疗问题(TeRP)、为查证医疗问题而进行检验(TeCP)以及医疗问题表明医疗问题(PIP).由于可供下载的部分仅包含 170 份训练集以及 256 份测试集,因此预处理阶段将其进行融合后按照 80∶20 的比例进行训练集与测试集的重新划分,对于同时包含多组实体关系的句子,为不同实体对构造该句作为新样本进行扩充,并将句中目标实体替换为相应实体类型[9],例如"He was given Lasix to prevent him from congestive heart failure."实体替换后变为:"He was given TREATMENT_A to prevent him from PROBLEM_B"(对他施与呋喃苯胺酸以防止患充血性心力衰竭).

SemEval 2013 Task-9 DDI:该语料库包含 Medline 生物医学文献的摘要部分以及 DrugBank 中由医师撰写的所有文档,共标注四种实体关系:建议(Advice),指同时用两种药的建议;影响(Effect),指药效的相互作用;机制(mechanism),指药物代谢动力学机制;以及无任何信息的药物交互(Int).原始数据集中包含 714 份训练集以及 191 份测试集,类似于 I2B2 处理方法,预处理阶段将同时包含多组关系的句子按照实体对进行样本扩充,句中的目标实体对按照前后顺序依次替换为 DRUG_A 与 DRUG_B,其余非目标实体均替换为 DRUG_N.这样的预处理技巧已在相关研究中得到证实[12,30,32].此外,早期研究[17,31-32]证明负样本的筛选有助于还原数据集实体的正确分布,因此本文沿用相同的负样本过滤方法,对于包含相同名称的目标实体对予以删除相应样本;如果实体对间属于别名关系,则删除该样本;若两目标实体在句中处于并列关系,则删除所属样本.表 2 所示的统计结果即为负样本过滤后的最终样本数.

参数设置:训练集上进行 5 折交叉验证,同时使用网格搜索(grid search)为两组数据集确定最佳模型参数,其中核心参数最优值参见表 3.实验中在 PubMed 开源数据集[33]上使用 GloVe[34]预训练词向量,作为词嵌入层的输入参与模型训练.

基线模型:选用 3 种类型共 6 个模型作为基准来比对本文所提出模型的优越性,其中包含现有研究中最佳性能模型 CRNN.基准模型详细信息如下:

表 3　关键参数

数据集	词嵌入维度	位置嵌入维度	通道数	Batchsize	Embedding层 dropout	Attention 层 dropout	LSTM units 数目	学习率	优化算法
I2B2	100	15	2	256	0.5	0.9	300	0.001	Adam
DDI	100	10	3	128	0.3	0.9	400	0.001	Adam

(1)机器学习模型:选用基于特征工程的 SVM 模型[14]作为该类别方法的代表,其人工特征包括句中各词与实体间的相对距离、词性(POS)以及实体标注(chunk tags)[10].该模型具体实现时选用 scikit-learn 库[34]中的 SVM 分类器.

(2)深度学习模型:分别选用 CNN-max[10]、Bi-LSTM[12]以及"recurrent+CNN"架构的 CRNN[11]进行实验比对.其模型参数均使用文献中所提到的最优值进行设置.三种基线模型隶属基础神经网络的简单应用,用于比对本文所提出的"recurrent+transformer"结构的优越性.其中,CRNN-Max 在现有研究中表现出最佳性能.

(3)注意力机制:CRNN-Att[11]与 LSTM-Att[12]是注意力机制在该任务上的成功应用.二者均将传统自注意力机制作为高层特征抽取器,对底层 RNN 或 CNN 的隐层输出做进一步特征抽取与编码.该类模型用于比对本文所提出的多通道自注意力机制的特征抽取能力.

实验环境:实验基于 Win10 操作系统,搭载 16 核 Intel(R) Xeon(R) E5-2623 V3 处理器与 128GB 运行时内存,核心计算力源于 4 块 NVIDIA TITAN X(Pascal) GPU 显卡,单显卡可用显存为 11.6GB.模型编码基于 Python3.6,采用 GLUON 深度学习框架构建模型.

4.2　模型整体性能对比

基线模型与本文所提出模型在两项数据集中的整体性能比对结果如表 4 所示,其中 nMCatt 表示含有 n 通道的注意力机制,两项数据集所对应的最佳通道数 n 参考表 5.WPos 表示引入基于权重的位置嵌入层,Wgt 指代带权学习的损失函数.表 4 中的 BLSTM-WPos-nMCatt-Wgt 为本文所使用的完整模型,其余包含 nMCatt、Wgt 或 WPos 标识的模型将作为本节细粒度实验比对,用于分别验证三项主要贡献点的有效性.

表 4 模型性能比对

类型	模型	2010 i2b2/VA				SemEval 2013 DDI			
		Precision/%	Recall/%	*F*1 score/%	Training Time/h	Precision/%	Recall/%	*F*1 score/%	Training Time/h
机器学习	SVM	69.81	56.09	60.28	1.59	65.39	40.13	49.74	0.71
基础神经网络模型	CNN-Max	42.17	74.63	52.61	1.21	68.15	46.58	54.05	0.44
	BLSTM-Max	57.54	55.40	55.60	1.37	73.98	59.96	65.41	0.50
	BLSTM-Att	65.23	56.77	60.04	1.44	53.43	64.86	58.27	0.51
	BLSTM-Wgt	65.61	57.22	61.19	1.24	71.23	60.59	65.69	0.49
recurrent＋CNN	CRNN-Max	71.91	62.09	65.21	1.98	72.91	60.88	65.89	0.74
	CRNN-Att	70.76	57.08	62.38	2.15	69.03	59.04	63.24	0.79
	CRNN-Max-WPos	69.86	61.59	65.46	1.97	72.08	61.35	66.28	0.75
	CRNN-Max-Wgt	70.34	63.74	66.87	1.86	73.52	63.51	68.14	0.69
recurrent＋transformer（本文模型及其变种）	BLSTM-nMCatt	66.61	63.67	65.09	1.49	72.21	66.30	68.94	0.54
	BLSTM-WPos-nMCatt	68.34	63.57	65.87	1.51	72.33	67.69	69.93	0.55
	BLSTM-nMCatt-Wgt	69.63	67.11	68.51	1.53	75.69	68.34	71.82	0.57
	BLSTM-WPos-nMCatt-Wgt	71.12	67.47	69.72	1.57	76.11	69.03	72.32	0.57

表 5 通道数对模型性能的影响

模型	通道数	2010 i2b2/VA			SemEval 2013 DDI		
		Precision/%	Recall/%	*F*1 score/%	Precision/%	Recall/%	*F*1 score/%
BLSTM -WPos-1MCatt-Wgt	1	68.74	66.03	67.36	73.12	68.67	70.82
BLSTM -WPos-2MCatt-Wgt	2	71.12	67.47	69.72	74.33	69.11	71.62
BLSTM -WPos-3MCatt-Wgt	3	71.01	67.93	69.43	76.11	69.03	72.32
BLSTM -WPos-4MCatt-Wgt	4	69.58	68.06	68.81	75.58	68.83	72.04
BLSTM -WPos-5MCatt-Wgt	5	68.12	69.01	68.56	73.18	67.66	70.31

在网络架构优越性方面，基于“recurrent＋transformer”架构的 BLSTM-WPos-nMCatt-Wgt 模型在两项数据集中的 *F*1 值（micro）均超越“recurrent＋CNN”架构的最优基线模型 CRNN-Max，其最高实现 6.43% 的性能提升（DDI 数据集）. 此外，在保证分类效果的基础上，“recurrent＋transformer”架构表现出较为明显的效率提升. 表 4 训练时间（Training Time）一栏统计了各模型达到最佳分类效果所需的训练时长，可以看出，在两数据集中，本文提出的 BLSTM-nMCatt（1.49 h，0.54 h）模型相较于 CRNN（2.15 h，0.79 h）最高实现 30.7% 与 31.6% 的效率提升，而完整模型 BLSTM-WPos-nMCatt-Wgt（1.57 h，0.57 h）也实现近 26% 与 27.8% 的提升. 该实验结果表明：相比于电子病历关系抽取领域主流的“recurrent＋CNN”架构，“recurrent＋transformer”可在保证下游关系抽取性能的基础上，进一步提升模型的训练效率，并降低模型整体的复杂度.

在模型有效性方面，仅施加多通道自注意力机制的 BLSTM-nMCatt 已逼近（I2B2）甚至超越（DDI）最佳基线模型 CRNN-Max，并且优于同样是“recurrent＋transformer”结构的 BLSTM-Att 模型. 该结果证实，相较于“recurrent＋CNN”为代表的最佳基线模型 CRNN-Max，“recurrent＋transformer”网络结构具有显著优越性，另一方面，本文所引入的多通道自注意力机制在电子病历数据集上的表现优于传统自注意力机制. 对于两项基于权重的辅助训练提升方法，表 4 中的 CRNN-Max-WPos、CRNN-Max-Wgt、BLSTM-WPos-nMCatt 以及 BLSTM-nMCatt-Wgt 分别验证了带权学习损失函数和基于权重的位置嵌入方法在不同网络架构中的有效性以及普适性. 同时，该实验结果可对比发现，施加于 CRNN-Max 中的两种辅助提升方法所带来的性能提升（*F*1 值最高提升 1.66%）远弱于 BLSTM-WPos-nMCatt 以及 BLSTM-nMCatt-Wgt（*F*1 值最高提升 3.30%），特别是带权学习的损失函数，其性能提升差异更为明显. 最终，完整模型 BLSTM-WPos-nMCatt-Wgt 相较于基线模型 BLSTM，在两数据集中的 *F*1 值最高实现 14.12% 的性能优化，相较于最佳基线模型，其性能提升近 4.51% 和 6.43%. 该对比实验结果表明，本文所提出的两种辅助训练提升方法具有一定的普适性，能够辅助提升模型对文本的表征能力，从而进一步影响下游关系抽取任务的效果，此外，实验结果指出，两种基于权重的辅助训练方法更加适用于基于“recurrent＋transformer”架构的模型.

上述对比结果表明，“recurrent＋transformer”

架构在电子病历数据集中表现出较大的性能提升以及较高的训练效率.对于前者,其原因在于以 transformer 架构为基础的多通道自注意力机制能够更大限度地捕捉电子病历复杂的实体分布特性以及交错关联的实体关系;而对于该架构所带来的训练效率提升,transformer 的提出者在文献[26]中分析了 Recurrent、CNN 及 transformer 模型单层网络结构的复杂度,其结论为 transformer 架构更具轻量级特点,由此带来训练效率的大幅提升.对比本文最优基线模型,CNN 卷积操作的时间复杂度同时受卷积核通道数、卷积核大小以及输出特征图大小的影响,呈现层内连乘、层间相加的形式,而多通道自注意力机制由于在各通道层面采取矩阵并行运算,因而其时间复杂度仅受文本序列长度影响而呈线性关系,同时,本文统计各对比模型训练时长,表 4 直观展示了本文模型在训练效率方面的优越性.此外,融入两项辅助训练提升方法的"recurrent + transformer"模型,一方面通过多通道自注意力机制对电子病历文本特征进行较好学习,另一方面结合使用带权学习的损失函数以降低参数空间的搜索范围,从而进一步实现分类效果的明显提升.

4.3　单项性能比较

以"recurrent + transformer"为基础架构的 BLSTM-MCatt 及其变种模型在整体性能上提升显著,为进一步验证围绕该架构所提出的三处主要贡献点的性能表现,本节将对其进行细粒度比对与分析.

(1)多通道自注意力机制

为了验证多通道自注意力机制的有效性,同时确定最佳通道数,表 5 列出了模型随通道数变化的性能表现.在两项数据集上,验证结果均呈现随通道数增加 $F1$ 值先增后减的趋势.这样的变化规律说明通道数的增加有助于模型对句子成分的深度挖掘,即能够进一步丰富语义层面的特征,使得模型参数空间更容易被拟合.然而,由于有限句长,通道数过多势必会引入部分冗余信息,从而干扰模型的正常训练,甚至在实验过程中观测到过拟合现象,因此会发现模型性能逐渐弱化的现象.

表 4 中 BLSTM-nMCatt 与 BLSTM-Att 的对比结果可以看出,多通道注意力机制的使用,在两项数据集中的 $F1$ 值分别提升近 5.05%和 10.67%,证实多通道自注意力机制对于电子病历的文本表征能力优于传统自注意力机制.值得注意的是,CRNN-Att 的性能表现优于 BLSTM-Att,同时本文所提出的 BLSTM-nMCatt 在两项数据集上却超越了 CRNN-Att.这一结果表明,在文本表征能力方面,本文所引入的多通道自注意力机制对句级别中局部和全局信息的学习能力优于 CNN,同时进一步验证了"recurrent + transformer"架构的优越性.

除上述直观对比外,对比表 4 中 BLSTM-Wgt、BLSTM-nMCatt 以及 BLSTM-nMCatt-Wgt 的 $F1$ 值可以发现,仅使用带权损失函数的 BLSTM-Wgt 模型其实验效果差于 BLSTM-nMCatt,而在施加多通道自注意力机制的基础上引入带权损失函数的 BLSTM-nMCatt-Wgt,却达到三者中的最优效果,且相较于 CRNN-Max 实现最大幅度的性能提升.这一现象表明,多通道自注意力机制对句级别语义信息具有较强的学习和捕捉能力,使得模型对电子病历文本有更好的表征性能,而在此基础上引入带权学习的损失函数可进一步提升模型在训练阶段的拟合能力,从而快速收敛至较优参数集.

(2)基于权重的位置嵌入

位置嵌入所携带的距离信息是 transformer 网络常用的文本特征表示方法,本文所提出的基于权重的位置嵌入可增强不同距离字符在 transformer 网络中的特征表达能力,有助于整体模型对具有复杂文本特性的电子病历进行表征与学习.

表 4 中所展示的 CRNN-Max-WPos 以及 BLSTM-WPos-nMCatt 是带有权重位置嵌入层的模型,在两项数据集上的验证结果表明,模型性能相较于 CRNN-Max 以及使用传统自注意力机制的 BLSTM-nMCatt 模型均有轻微提升.该对比结果表明:一方面基于权重的位置嵌入方法在各架构中具有一定的普适性;而另一方面说明,针对实体高密度分布,且相同实体交错参与构成不同关系对的电子病历文本,含有权重的位置嵌入法能够进一步放大位置信息的重要性,即降低无用句成分信息的权重,提高关键特征的重要性,以此扩大句成分间重要性的差异度,从而提升模型对文本的表征能力.此外,基于权重的位置嵌入法之所以仅表现出微弱的提升,其原因在于较短的有限句长(I2B2 最大句长为 204,而 DDI 为 144)使得位置信息间较小的区分度不足以为模型训练贡献更多的有效特征.

(3)带权学习的损失函数

多通道自注意力机制单项实验中,通过对比 BLSTM-Wgt、BLSTM-nMCatt 以及 BLSTM-nMCatt-Wgt 三项模型,证实了句级别表征能力更强的多通道自注意力机制有助于带权损失函数发挥其更大功效.因此,本文将带权损失函数作为模型的辅助训练

提升方法,进一步分析其在不同网络架构中的普适性以及细粒度分类性能的有效性.

带权学习的损失函数在单项比对中带来较为明显的性能提升.对于"recurrent+transformer"架构,相比于 BLSTM-nMCatt,BLSTM-nMCatt-Wgt 在两项数据集中 $F1$ 值分别提升近 3.8%和 2.9%,并在 3 项评价指标上超越基线最优模型 CRNN-Max.此外,在"recurrent+CNN"架构中引入该辅助训练方法的 CRNN-Max-Wgt 模型,相较于 CRNN-Max,同样实现了较为可观的性能提升(1.66%).该对比结果表明,带权学习的损失函数具有一定的普适性,且在不同架构中均具有辅助训练的提升作用,特别是与本文的 transformer 架构网络相结合的多通道自注意力机制,其提升效果更为显著.

为了进一步观测引入带权学习损失函数的"recurrent+transformer"架构对类别不均衡数据的拟合能力,表 6 详细展示了单类别细粒度性能表现,结合表 1 与表 2 对两项数据集类别分布的统计情况,可以直观发现:BLSTM-nMCatt-Wgt 对于小类别样本的分类能力相较于基线模型有显著提升,例如 TrWP、TrIP、TrNAP 以及 Int,其中在 TrNAP 类上实现 23.55%的最大提升.对于多数类样本(TeRP、TrAP、Advice),模型同样实现较为可观的性能提升.在中等规模类别 PIP 上的验证结果弱于基线模型 BLSTM-Att,其原因在于,该类关系所包含的两项实体为同类,相较于其他类型样本,多通道自注意力机制将因此代入较多冗余信息,从而影响模型在该类别的拟合能力.此外,实验过程中发现,PIP 类的 $F1$ 指标随训练轮数的增加先升后降,同样验证了因冗余信息而导致的模型过拟合现象.

表 6 细粒度类别性能对比

关系	BLSTM-Max	BLSTM-Att	CRNN-Max	CRNN-Att	BLSTM-nMCatt	BLSTM-WPos-nMCatt	BLSTM-nMCatt-Wgt	BLSTM-WPos-nMCatt-Wgt
TrCP	35.48	40.01	43.18	47.66	55.12	55.97	57.01	58.51
TrAP	63.40	61.38	67.39	63.94	71.01	72.87	74.10	74.37
TrWP	0.00	0.00	16.67	9.52	15.33	12.39	21.13	19.02
TrIP	0.00	13.33	25.71	34.48	5.89	9.82	39.21	40.39
TrNAP	0.00	10.29	36.36	18.60	16.79	17.79	58.33	59.91
TeRP	79.50	81.12	80.32	76.31	81.37	82.36	81.91	82.59
TeCP	21.20	38.36	39.46	39.76	52.88	53.90	57.11	59.07
PIP	56.05	61.00	58.04	55.53	51.87	52.01	55.13	57.94
Advice	68.31	62.43	70.12	63.13	74.26	75.76	78.76	81.14
Mechanism	67.13	61.89	66.87	67.04	71.32	74.21	73.72	74.06
Effect	66.30	58.31	65.93	64.22	67.07	68.31	69.21	69.48
Int	49.45	42.56	52.10	51.32	49.94	50.35	53.73	53.23

损失函数的优劣直接影响模型对参数空间的拟合效果,无论是粗粒度的整体比对还是细粒度的类别对比,都证实引入带权学习的损失函数提升了模型对参数空间的拟合能力.此外,实验过程中发现,类别权重的损失函数在保证模型性能的同时,加速了训练过程的收敛速率.图 2 展示了类别权重对模型收敛速度的影响情况,该实验基础参数设置基于表 2 所示的最佳参数,仅修改损失函数类型进行对比实验.两数据集上分别进行 10 次实验,每 10 轮记录一次当前测试集的 $F1$ 值,并通过设定早停机制(early stop)来中断训练.对 10 次结果中每轮 $F1$ 取均值作为该轮真实 $F1$ 值.从图 2 中可以看出,DDI 数据集上,带权损失函数在 50 到 60 轮之间便可达到最优,而使用普通交叉熵损失函数时,则需 70 轮左右模型到达收敛.I2B2 数据集上表现出同样的规律.这一验证结果表明,本文所提出的基于类别权重的交叉熵损失函数在带来模型性能提升的同时,加速了模型收敛速率.其原因在于各类别对总体损失的贡献得到平衡,因此一定程度上缩小了模型的参数搜索空间,从而提升模型的收敛速率.

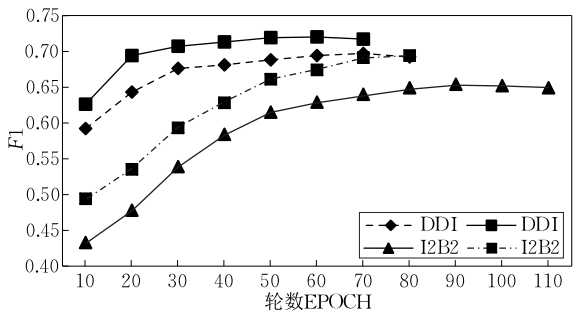

图 2 带权损失函数对模型收敛速率的影响

4.4 注意力效果案例分析

图 3 所示的热力图展示了多通道注意力机制对句级别特征的学习效果.该案例来源于 I2B2 语料集,实体间关系为 TrWP,预处理时将目标实体替换为相应实体类型,如图 3 中的 PROBLEM_B 以及

TREAMENT_A. 图 3(1)～图 3(4)分别展示了通道数对句级别语义特征的学习效果，颜色越深表示该字符具有较高的注意力权重，反之代表较低的重要性. 通道数为 1 可认为与传统自注意力机制等价，相比于 2 通道的注意力效果，传统自注意力机制误将“improved”给予较高权重，并且未能较好地对该例关系判别起重要作用的“ but”和“recurred”赋予更高的注意力权重. 该对比结果证实了多通道自注意力机制能够更好地捕捉句中的关键语法成分，热力图中可直观看到，各句子成分被赋予差别更大的重要度，进一步提升有用句成分信息对文本表征的能力，从而保证模型在电子病历关系抽取中的性能. 随着通道数的增加，字符间重要性的区分度有所下降，结合表 5 的比对结果可知，通道数过高会导致冗余信息的引入，从而使得模型出现过拟合现象. 总之，该例的注意力权重可视化结果表明，句级别语义特征能够在多通道注意力机制的作用下得到更好的捕捉和表达，如案例中的“recurred on”以及连接词“and”.

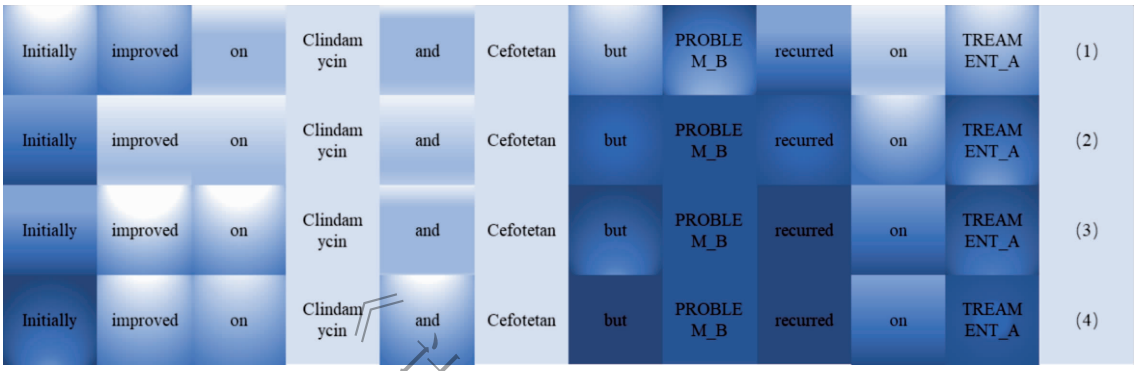

图 3　多通道注意力权重热力图

5　结　论

本文提出并验证了“recurrent＋transformer”网络架构在电子病历实体关系抽取任务中的有效性，首次引入多通道自注意力机制作为 transformer 层，并围绕上述架构提出两种基于权重的辅助训练提升方法，以提升模型对电子病历专属文本特点的表征与学习能力. 模型整体主要包含三层核心网络结构：(1) 多通道自注意力层通过拟合多组权重向量来表达句子各成分间重要性，以此捕捉句级别语义信息，采用矩阵操作实现多通道注意力的并行计算，以代入惩罚项的方式来尽可能避免对冗余信息的过多学习. 单项对比验证及注意力权重可视化结果表明，相比于传统自注意力机制，多通道注意力机制能够有效捕捉实体密度高、实体间关系复杂等文本特征. (2) 针对医疗领域数据分布的独特性，首次在该任务提出类别权重计算方法并与损失函数同时参与训练，以此实现对数据真实分布的拟合，同时保证模型在各类别中的学习能力. 该损失函数显著提升模型对小类别样本的拟合能力，同时加速模型的收敛速率. (3) 位置嵌入与注意力机制的搭配使用已在各领域得到证实，针对电子病历高密度的实体分布特性，本文为字符间相对位置赋予权重，以此得到的位置嵌入层与传统词嵌入层进行向量拼接，作为文本信息的初级表征. 该嵌入层计算方法实现了对位置信息的有效放缩，提升了距离特征在模型训练过程中的区分度.

经过粗细粒度的实验验证与比对，本文所提出的模型及其相关变种在 2010 I2B2/VA 及 SemEval 2013 DDI 数据集中均表现出较好的优越性. 其中 BLSTM-WPos-nMCatt-Wgt 在三项评价指标上均取得最佳效果，相较于最优基线模型 CRNN-Max，$F1$ 指标分别表现出 4.51％和 6.43％的明显提升. 其中，在单项比对中，多通道注意力机制的使用和基于类别权重的损失函数带来最为明显的性能提升，该实验效果表明，本文所提出的方法能够实现对电子病历文本特殊性的掌控，无论是句级别语义特征的挖掘还是数据集真实分布的拟合都实现较好的匹配与迎合. 此外，仅使用 BiLSTM 与多通道注意力机制相结合的 BLSTM-nMCatt 模型在性能表现上超越 CRNN-Att，这一验证结果表明，模型复杂度更低的“recurrent＋transformer”结构在该任务中同样能够实现较好的性能. 本文探索的方法正是为进一步提升模型对文本特征的表征能力，从而提升实体关系抽取在电子病历上的性能，这一切入点也是当下的主流思路之一. 在未来工作中，我们将进一步探索本文方法的横向扩展性能，并探索高密度实体

分布及实体关系复杂多变领域文本的通用解决方案.此外,相比于通识领域,电子病历文本仍存在很多独特差异性(如中医的文言式句法),因此,电子病历文本特征的表达能力仍是有较大的研究与提升空间,这也是我们将在未来亟待研究和解决的问题之一.

参 考 文 献

[1] Wasserman R C. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research. Academic Pediatrics, 2011, 11(4): 280-287

[2] Mooney R J, Bunescu R C. Subsequence kernels for relation extraction//Proceedings of the advances in Neural Information Processing Systems. Hong Kong, China, 2006: 171-178

[3] Yang Jin-Feng, Guan Yi, He Bin, et al. Corpus construction for named entities and entityrelations on Chinese electronic medical records. Journal of Software, 2016, 27(11): 2725-2746(in Chinese)
(杨锦锋,关毅,何彬等.中文电子病历命名实体和实体关系语料库构建.软件学报,2016,27(11):2725-2746)

[4] Yang Jin-Feng, Yu Qiu-Bin, Guan Yi, et al. An overview of research on electronic medical record oriented named entity recognition andentity relation extraction. Acta Automatica Sinica, 2014,40(8): 1537-1562(in Chinese)
(杨锦锋,于秋滨,关毅等.电子病历命名实体识别和实体关系抽取研究综述.自动化学报,2014,40(8):1537-1562)

[5] Rink B, Harabagiu S. Utd: Classifying semantic relations by combining lexical and semantic resources//Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden, 2010: 256-259

[6] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2. Singapore, 2009: 1003-1011

[7] Bunescu R C, Mooney R J. A shortest path dependency kernel for relation extraction//Proceedings of the Conference on Human Language Technology and Empirical methods in Natural Language Processing. British Columbia, Canada, 2005: 724-731

[8] Xu Y, Mou L, Li G, et al. Classifying relations via long short term memory networks along shortest dependency paths//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 1785-1794

[9] Cai R, Zhang X, Wang H. Bidirectional recurrent convolutional neural network for relation classification//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany, 2016: 756-765

[10] Sahu S K, Anand A, Oruganty K, et al. Relation extraction from clinical texts using domain invariant convolutional neural network. arXiv preprint arXiv: 1606.09370, 2016

[11] Raj D, SAHU S, Anand A. Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text//Proceedings of the 21st Conference on Computational Natural Language Learning. Vancouver, Canada, 2017: 311-321

[12] Sahu S K, Anand A. Drug-drug interaction extraction from biomedical texts using long short-term memory network. Journal of Biomedical Informatics, 2018, 86(5): 15-24

[13] He B, Guan Y, Dai R. Convolutional gated recurrent units for medical relation classification//Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine. Madrid, Spain, 2018: 646-650

[14] Li P, Mao K, Yang X, et al. Improving Relation Extraction with Knowledge-attention. arXiv preprint arXiv: 1910.02724, 2019

[15] Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. Journal of the American Medical Informatics Association, 2011, 18(5): 594-600

[16] Chowdhury M F M, Lavelli A. FBK-irst: A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information//Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Atlanta, USA, 2013: 351-355

[17] Kim S, Liu H, Yeganova L, et al. Extracting drug—drug interactions from literature using a rich feature-based linear kernel approach. Journal of Biomedical Informatics, 2015, 55(2): 23-30

[18] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea, 2012: 1201-1211

[19] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network//Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland, 2014: 2335-2344

[20] Fu L, Nguyen T H, Min B, et al. Domain adaptation for relation extraction with domain adversarial neural network//Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Taipei, China, 2017: 425-429

[21] Lee J Y, Dernoncourt F, Szolovits P. Mit at semeval-2017 task 10: Relation extraction with convolutional neural networks. arXiv preprint arXiv: 1704.01523, 2017

[22] Peng N, Poon H, Quirk C, et al. Cross-sentence n-ary relation extraction with graph lstms. Transactions of the Association for Computational Linguistics, 2017, 5: 101-115

[23] Zhang Y, Qi P, Manning C D. Graph convolution over pruned dependency trees improves relation extraction. arXiv preprint arXiv:1809.10185, 2018

[24] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 207-212

[25] He B, Guan Y, Dai R. Classifying medical relations in clinical text via convolutional neural networks. Artificial Intelligence in Medicine, 2019, 93: 43-49

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. California, USA, 2017: 5998-6008

[27] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[28] Lin Z, Feng M, Santos C N, et al. A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130, 2017

[29] Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association, 2013, 20(5): 806-813

[30] Rastegar-Mojarad M, Boyce R D, Prasad R. UWM-TRIADS: Classifying drug-drug interactions with two-stage SVM and post-processing//Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Atlanta, USA, 2013: 667-674

[31] Liu S, Tang B, Chen Q, et al. Drug-drug interaction extraction via convolutional neural networks. Computational and Mathematical Methods in Medicine, 2016, 2016: 1-8

[32] Zhao Z, Yang Z, Luo L, et al. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics, 2016, 32(22): 3444-3453

[33] Sahu S, Anand A. Evaluating distributed word representations for capturing semantics of biomedical concepts//Proceedings of the BioNLP 15. Beijing, China, 2015: 158-163

[34] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2011, 12(10): 2825-2830

NING Shang-Ming, M. S. candidate. His main research interests focus on natural language processing.

TENG Fei, Ph. D., associate professor. Her main research interests focus on cloud computing and data mining.

LI Tian-Rui, Ph. D., professor. His main research interests focus on artificial intelligence, data mining, granular computing and rough sets, cloud computing and big data.

Background

Relation classification is an important task in natural language processing, which is applied in clinical records recent years to extract valuable medical or clinical information, and assist in building the upper application like knowledge graph. There have been many researches in general domain, but it is a challenge task in clinical text. The biggest difficulty in medical record text is that there are multiple different types of entities appear in the same sentence at the same time, and these entities could be combined into different relations. Thus, some researches have not solved that very well, and they think of building joint model with RNNs and CNNs is a good way, however, the result still does not meet expectations.

Based on previous works, we pay more attention to sentence-level features to try to solve the problems above, and propose a "recurrent+transformer" architecture with multi-channel self-attention mechanism to enrich the semantic features of the sentence level. Experimental results show that, our model is more efficient than "recurrent+CNN", and the highest performance improvement is 10.67%, compared to traditional self-attention in $F1$ score. Our model indicates that a "recurrent+transformer" idea is also effective but with a simpler network structure, and it gives us a new direction to continue to improve networks.