基于双路分段注意力神经张量网络的临床文本关系抽取

隗 昊[1,2],唐焕玲[3],周 爱[2],张益嘉[2],陈 飞[2],鲁明羽[2]
(1. 大连外国语大学软件学院,辽宁大连 116044; 2. 大连海事大学信息科学技术学院,辽宁大连 116026;
3. 山东工商学院计算机科学与技术学院,山东烟台 264005)

摘 要: 目前,生物医学领域的关系提取工作已经取得了长足的发展,但是在面对句式复杂的临床医学文本时,由于存在大量长句以及句中实体对的高密度分布,限制了当前关系抽取模型性能的进一步提升. 本文提出了一种基于张量权重矩阵的双向门控循环单元网络(Tensor-based Bidirectional Gated Recurrent Unit, Tensor-BiGRU)和分段注意力机制的关系抽取模型,基于张量权重矩阵改进BiGRU网络的编码方式,提升神经网络捕获底层特征的能力,而后提出了两种分段注意力机制,以提高模型捕获长句特征的性能. 此外,当句子中有多个实体对时,引入实体对的语义信息特征来克服模型的性能下降. 本文进一步提出一种权重自适应的交叉熵损失函数,用于提升模型面对数据集中不同关系类别的样本分布不平衡问题的泛化性. 实验结果表明,在不依赖任何特征工程和高性能运算环境的情况下,本文模型在2010 i2b2/VA临床关系抽取数据集上实现了先进的性能.

关键词: 关系抽取;临床文本;神经张量网络;分段注意力机制;样本不平衡
基金项目: 国家自然科学基金(No.61976124, No.62072070)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2023)03-0658-08
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210628

Clinical Relation Extraction via Dual Piecewise Attention Neural Tensor Network

WEI Hao[1,2], TANG Huan-ling[3], ZHOU Ai[2], ZHANG Yi-jia[2], CHEN Fei[2], LU Ming-yu[2]
(1. *School of Software*, *Dalian University of Foreign Languages*, *Dalian*, *Liaoning* 116044, *China*;
2. *Information Science and Technology College*, *Dalian Maritime University*, *Dalian*, *Liaoning* 116026, *China*;
3. *School of Computer Science and Technology*, *Shandong Technology and Business University*, *Yantai*, *Shandong* 264005, *China*)

Abstract: At present, biomedical relation extraction has made considerable progress. However, when dealt with complex clinical texts, due to the large number of long sentences and the high density distribution of entity pairs in the sentences, the existing methods of relation extraction still have defects. We propose a relation extraction model via tensor-based bidirectional gate recurrent unit (Tensor-BiGRU) and piecewise attention mechanism. The ability of BiGRU to extract the underlying features is enhanced based on tensor weight matrix. Two kinds of piecewise attention mechanisms are proposed to improve the performance of the model in capturing long sentence features.When the sentence has multiple entity pairs, the semantic representations of the entity pairs are introduced to overcome the performance degradation of the mode. A weight-adaptive cross-entropy loss function is proposed to improve the generalization of the model when the sample distribution of different relation categories in the dataset is unbalanced. The experimental results show that without relying on any feature engineering and high-performance computing environment, the model achieves advanced performance on the 2010 i2b2/VA clinical data set.

Key words: relation extraction; clinical texts; neural tensor network; piecewise attention mechanism; sample imbalance
Foundation Item(s): National Natural Science Foundation of China (No.61976124, No.62072070)

收稿日期:2021-05-18;修回日期:2021-07-16;责任编辑:梅志强

1　引言

关系抽取是自然语言处理的关键和基本任务之一,旨在从大规模非结构化或半结构化的自然语言文本中发现并抽取实体之间存在的某种预定义的语义关系[1,2]. 在生物医学领域,临床文本关系抽取被用来挖掘以电子病历记录为主要载体的临床实体(如药物、疾病、症状等)间存在的某种关系[3]. 如图1所示,该例句存在四个临床实体,共产生两种关系类型的四个关系三元组. 面向临床文本的关系抽取对医疗图谱构建、药物重定位、在线诊疗系统开发等具有重要的现实意义.

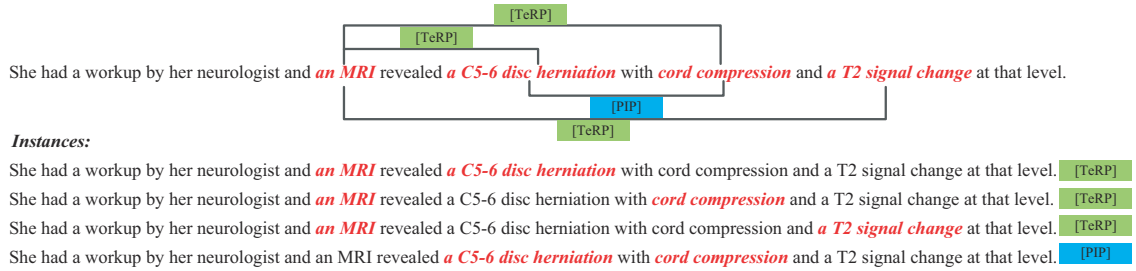

图1　临床文本关系抽取示例

临床文本关系抽取的相关研究大致可分为基于传统的机器学习方法和基于深度学习方法. 机器学习方法大多都基于丰富特征和核函数实现,Rink等人[4]基于上下文信息、实体类型信息、嵌套关系信息等构建了丰富的特征表示,并使用支持向量机进行关系抽取并取得了较好的性能. De Bruijn等人[5]充分利用了领域知识构建特征,并引入了海量的未标注数据基于自训练的半监督学习策略构建了最大熵关系抽取模型,实验证明了增加训练数据量和引入外部知识的有效性. 以神经网络模型为核心的深度学习方法近年来被广泛应用于生物医学领域的临床文本关系抽取工作. Sahu等人[6]利用词向量、位置向量、词性向量、词块向量和实体类型向量作为输入的特征表示并构建了卷积神经网络(Convolutional Neural Networks, CNN)提取临床文本中的关系. Luo等人[7]将长短时记忆网络(Long Short-Term Memory, LSTM)用于提取临床文本关系,根据目标实体对的位置构造了一个分段LSTM模型,分别提取目标实体周围短程的语义信息,在一定程度上缓解了远距离实体对关系提取问题. He等人[8]提出一个多池化操作的CNN模型,用于增加样本区分度以解决句内关系样本间的输入特征近似问题,并设计了类别约束的损失函数用于缓解类别不平衡问题. Li等人[9]设计了一个分段注意力BiLSTM(Bidirectional LSTM)模型,用于捕获复杂的临床文本长句,并通过训练一个特征矩阵学习实体对之间的语义关系. 宁尚明等人[10]以注意力机制为核心,提出一个基于“recurrent+ transformer”架构的关系抽取模型,其中多通道自注意力机制可以捕获更丰富的句子级语义信息,同时设计了两种辅助训练策略,在优化小样本类别拟合能力的同时提升模型的训练效率.

先前研究表明,与通用领域的关系抽取语料相比,临床文本语料在平均序列长度更短的前提下所包含的实体数量为通用领域的2到3倍,存在关系的实体对更为通用领域4到6倍[10]. 如图1所示,完整例句为一个由26个单词组成的长句,句式结构复杂且实体分布集中,每个实体之间仅间隔一个单词,因此实体对周围的上下文所包含的语义信息较为混乱. 另外,真实情况下的临床文本语料存在样本分布不平衡问题,现有的部分模型在学习小样本类别特征时存在较大的误差. 针对上述归纳的当前临床医学文本句式结构冗长复杂且关系实体对分布密集的特点,本文提出一种基于Tensor-BiGRU和分段注意力机制的临床文本关系抽取模型,我们的主要贡献总结如下:

(1) 基于张量权重矩阵改进BiGRU网络的编码方式,用于更充分地学习和提取输入表示的底层语义特征.

(2) 设计了分段注意力层并引入实体对语义表示层用于提升模型对结构冗长复杂且实体对分布密集的临床医学长句的特征捕获能力.

(3) 提出了权重自适应的交叉熵损失函数用于缓解临床医学数据集中样本不平衡问题.

2　模型介绍

图2为本文所提出的神经网络关系抽取模型的整体框架,由六部分构成:输入层、嵌入层、Tensor-BiGRU层、分段注意力层、实体对表示层和输出层.

2.1　嵌入层

本文使用Word2Vec[11]训练工具生成词向量. 设样本句子X由n个词组成,则$X=\{t_1,t_2,t_3,\cdots,t_n\}$,令$\boldsymbol{t}_i$为$t$的第$i$个词的one-hot表示形式,词向量$\boldsymbol{x}_i$表示为:

$$\boldsymbol{x}_i=\boldsymbol{W}^{\text{emb}}\boldsymbol{t}_i \tag{1}$$

其中$\boldsymbol{W}^{\text{emb}}\in\mathbf{R}^{d\times|v|}$为训练得到的词向量表;$\boldsymbol{t}_i\in\mathbf{R}^{|v|}$,$\boldsymbol{x}_i\in\mathbf{R}_d$,$d$为预设词向量维度,$|v|$为one-hot形式的词表

大小. 另外,本文引入了位置[12]信息 $\boldsymbol{W}^{\mathrm{p}} \in \mathbf{R}^{2d_{\mathrm{p}} \times |v|}$ 和实体类型[6]信息 $\boldsymbol{W}^{\mathrm{et}} \in \mathbf{R}^{d_{\mathrm{et}} \times |v|}$ 两种辅助特征作为输入表示. 两种辅助特征以随机初始化的方式生成,采用拼接的方式将三种特征向量串联后生成最终的输入特征表示 $\boldsymbol{W}$,其中 d_{p} 为预设位置向量的维度,d_{et} 为预设实体类型向量的维度.

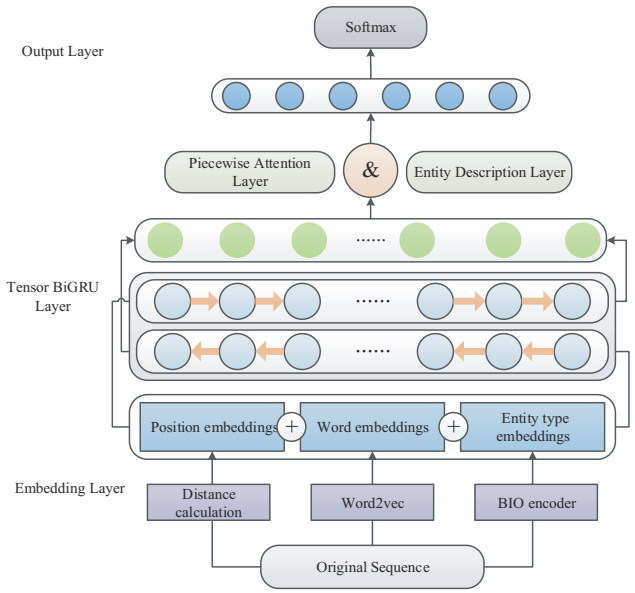

图2 基于Tensor-BiGRU和分段注意力机制的临床文本关系抽取模型框架

2.2 Tensor-BiGRU层

临床医学文本通常句式复杂且关系实体对分布密集,这大大影响了编码层的特征提取效果. 标准的循环神经网络(Recurrent Neural Network, RNN)及其变种结构中输入表示仅通过与非线性函数的隐式交互进行特征提取,为了提升编码层捕获特征的能力,受启发于RNTN(Recursive Neural Tensor Network)[13],本文在编码层提出一种基于张量权重矩阵的BiGRU网络—Tensor-BiGRU,通过初始化一个张量权重矩阵进行乘积运算以生成隐藏层的候选特征输出,这样可以使BiGRU网络的隐藏层和输入特征向量之间实现更充分地交互. Tensor-BiGRU首先通过两种门控结构选择性的过滤和保存上下文全局信息,而当前时刻的候选输出由上一时刻的隐藏层输出和当前时刻的输入共同决定,相关计算公式如下:

$$\boldsymbol{z}_t = \sigma\left(\boldsymbol{w}_z\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_z\right) \tag{2}$$

$$\boldsymbol{r}_t = \sigma\left(\boldsymbol{w}_r\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_r\right) \tag{3}$$

$$\tilde{\boldsymbol{h}}_t = \tanh\left(\left[\boldsymbol{r}_t * \boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right]\boldsymbol{w}_{\mathrm{T}}\begin{bmatrix}\boldsymbol{r}_t * \boldsymbol{h}_{t-1} \\ \boldsymbol{x}_t\end{bmatrix} + \boldsymbol{w}\left[\boldsymbol{r}_t * \boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_h\right) \tag{4}$$

$$\boldsymbol{h}_t = \left(1 - \boldsymbol{z}_t\right) * \boldsymbol{h}_{t-1} + \boldsymbol{z}_t * \tilde{\boldsymbol{h}}_t \tag{5}$$

其中,$\left[\boldsymbol{r}_t * \boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right]\boldsymbol{w}_{\mathrm{T}}\begin{bmatrix}\boldsymbol{r}_t * \boldsymbol{h}_{t-1} \\ \boldsymbol{x}_t\end{bmatrix}$ 为张量权重矩阵的乘积运算,$\boldsymbol{x}_t$ 为当前时刻的输入,$\boldsymbol{r}_t$ 和 $\boldsymbol{z}_t$ 分别为重置门和更新门的输出,$\boldsymbol{h}_{t-1}$ 为上一时刻隐藏层的输出,$\boldsymbol{w}_{\mathrm{T}}$ 为初始化的张量权重矩阵,$\boldsymbol{w}_{\mathrm{T}} \in \mathbf{R}^{d_{W+\mathrm{TG}} \times d_{W+\mathrm{TG}} \times d_{\mathrm{TG}}}$,$d_{\mathrm{TG}}$ 为Tensor-BiGRU中隐藏层神经单元的维度,$\boldsymbol{w}$、$\boldsymbol{w}_r$ 和 $\boldsymbol{w}_z$ 为参与训练的参数矩阵. 另外,我们将标准GRU网络的隐藏层生成的候选特征输出 $\boldsymbol{w}\left[\boldsymbol{r}_t * \boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right]$ 与基于张量矩阵计算后的输出相加,生成最终的候选特征输出 $\tilde{\boldsymbol{h}}_t$. $\tilde{\boldsymbol{h}}_t$ 的计算过程如图3所示,其中左框内为Tensor-BiGRU的隐藏层候选输出,右框内为BiGRU的隐藏层候选输出,绿色圆点分别表示 $\boldsymbol{w}_r$ 和 $\boldsymbol{w}$,蓝色和黄色圆点分别表示 $\boldsymbol{r}_t * \boldsymbol{h}_{t-1}$ 和 $\boldsymbol{x}_t$.

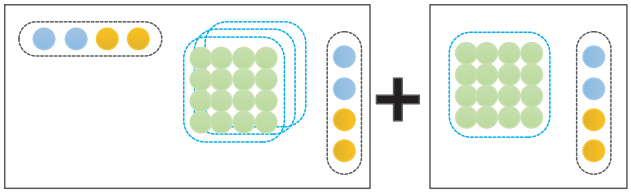

图3 Tensor-BiGRU隐藏层和输入特征交互示意图

2.3 分段注意力层

2.3.1 分段知识感知注意力

当前基于注意力机制的神经网络模型已经被广泛地用于多种自然语言处理任务. 通常,在RNN类网络结构中使用的注意力机制大都基于原始序列自身生成权重矩阵,与外部特征无关[14]. 但是面对句式冗长、实体对分布密集的临床医学文本时,在输入原始序列特征表示的同时充分地利用额外输入特征可以更准确地表达其语义信息. 因此,本文提出了一种分段知识感知的注意力机制,该机制在分段处理复杂长句的同时充分考虑了全局位置信息以及实体类型信息.

如图4的(a)部分所示. 首先,将Tensor-BiGRU模型的输出按目标实体对的位置划分为五段,分别为目标实体对和首实体前序列、实体对中间序列、尾实体后序列,然后将目标实体对分别与三段序列拼接后输入知识感知注意力模块. 令 $\boldsymbol{H}_{1/2/3}$ 为划分后的Tensor-BiGRU模型的三段输出,注意力权重及重分配后的输出表示的计算公式如下:

$$\boldsymbol{H}_{1/2/3} = \left[\vec{\boldsymbol{h}}_t \oplus \overleftarrow{\boldsymbol{h}}_t\right] \tag{6}$$

$$\boldsymbol{M}_{\mathrm{p/e}} = \tanh\begin{bmatrix}\boldsymbol{w}_h \boldsymbol{H}_{1/2/3} \\ \boldsymbol{w}_{\mathrm{p/e}} \boldsymbol{H}_{1/2/3}\end{bmatrix} \tag{7}$$

$$\boldsymbol{\alpha}_{\mathrm{p/e}} = \mathrm{softmax}\left(\boldsymbol{w}_{\mathrm{mp/me}} \boldsymbol{M}_{\mathrm{p/e}}\right) \tag{8}$$

$$\boldsymbol{O}_{\mathrm{p/e}} = \boldsymbol{H}_{1/2/3} \boldsymbol{\alpha}_{\mathrm{p/e}} \tag{9}$$

其中,$\boldsymbol{H} \in \mathbf{R}^{2d \times n_{1/2/3}}$,$\boldsymbol{M}_{\mathrm{p}} \in \mathbf{R}^{2d+2d_{\mathrm{p}} \times n_{1/2/3}}$ 为位置感知注意力

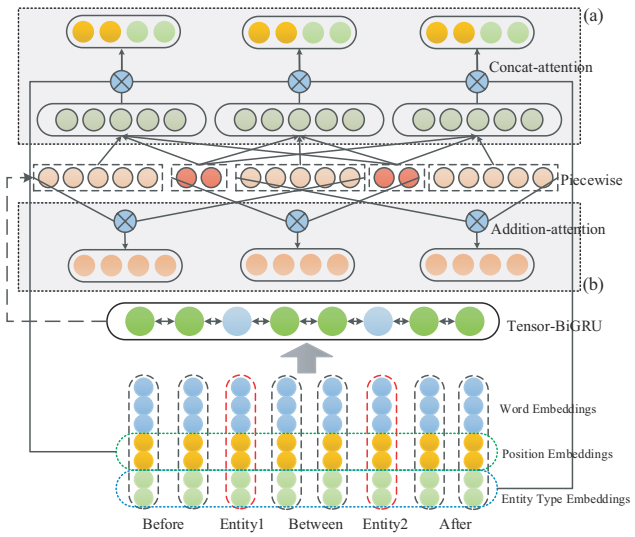

图4　分段注意力层内部结构图

输入，$\boldsymbol{M}_{\mathrm{e}} \in \mathbf{R}^{2d+d_{\mathrm{et}} \times n_{1/2/3}}$ 为实体类型感知注意力输入．$\boldsymbol{w}_h \in \mathbf{R}^{2d \times 2d}$、$\boldsymbol{w}_{\mathrm{p}} \in \mathbf{R}^{2d_{\mathrm{p}} \times 2d_{\mathrm{p}}}$、$\boldsymbol{w}_{\mathrm{e}} \in \mathbf{R}^{d_{\mathrm{et}} \times d_{\mathrm{et}}}$、$\boldsymbol{w}_{\mathrm{mp}} \in \mathbf{R}^{2d+2d_{\mathrm{p}}}$ 和 $\boldsymbol{w}_{\mathrm{me}} \in \mathbf{R}^{2d+d_{\mathrm{et}}}$ 均为训练期间不断更新的参数矩阵．$\boldsymbol{\alpha}_{\mathrm{p/e}} \in \mathbf{R}^{n_{1/2/3}}$ 表示维度与输入序列中的特征向量个数一致的注意力权重矩阵．$\boldsymbol{O}_{\mathrm{p/e}} \in \mathbf{R}^{2d}$ 表示特征向量加权求和后的最终输出．

2.3.2　分段注意力池化

受到文献[12]的启发，本层设计了分段注意力池化模块用于进一步捕获复杂临床文本长句中的重要特征，如图4的(b)部分所示．与上一节类似，该模块首先将Tensor-BiGRU层的输出切分为“序列1+实体1”、“实体1+序列2+实体2”和“实体2+序列3”三部分，然后将切分后的三段特征向量分别进行加性注意力计算生成三组权重矩阵 $\boldsymbol{\alpha}_1$、$\boldsymbol{\alpha}_2$ 和 $\boldsymbol{\alpha}_3$，向量中的值即为分配给输入序列中每个特征向量的权重．将三段特征向量与注意力权重加权求和后拼接构成分段注意力池化模块的输出．具体公式如下：

$$\boldsymbol{M}_{1/2/3} = \tanh\left(\boldsymbol{H}_{1/2/3}\right) \tag{10}$$

$$\boldsymbol{\alpha}_{1/2/3} = \mathrm{softmax}\left(\boldsymbol{w}\boldsymbol{M}_{1/2/3}\right) \tag{11}$$

$$\boldsymbol{O}_{1/2/3} = \boldsymbol{H}_{1/2/3}\boldsymbol{\alpha}_{1/2/3} \tag{12}$$

其中，$\boldsymbol{H}_{1/2/3} \in \mathbf{R}^{2d \times n_{1/2/3}}$，$\boldsymbol{w} \in \mathbf{R}^{2d}$ 为训练参数矩阵，$\boldsymbol{\alpha}_{1/2/3} \in \mathbf{R}^{n_{1/2/3}}$ 为注意力权重矩阵，$\boldsymbol{O}_{1/2/3} \in \mathbf{R}^{2d}$ 为模块的最终输出．

2.4　实体对语义表示层

从临床医学语料抽取关系的难点之一在于样本序列中存在关系的实体对的密集分布．已有研究表明，样本序列中实体对分布越密集，对目标实体对抽取关系时的难度越大．针对上述问题，我们引入了实体对的语义信息特征表示[9]，用于学习目标实体对之间的交互作用．设Tensor-BiGRU编码后的目标实体1和2分别为 $\boldsymbol{t}_1$ 和 $\boldsymbol{t}_2$．由于不同序列中的实体长度不一，因此首先将他们转换为固定维度的向量，然后将其拼接并通过两种权重矩阵得到实体对的语义特征表示 $\boldsymbol{t}$，相关计算公式如下：

$$\boldsymbol{m}_{1/2} = \tanh\left(\boldsymbol{t}_{1/2}\right) \tag{13}$$

$$\boldsymbol{\alpha}_{1/2} = \mathrm{softmax}\left(\boldsymbol{w}\boldsymbol{m}_{1/2}\right) \tag{14}$$

$$\boldsymbol{E}_{1/2} = \boldsymbol{m}_{1/2}\boldsymbol{\alpha}_{1/2} \tag{15}$$

$$\boldsymbol{t}_{\mathrm{ep1}} = \begin{bmatrix}\boldsymbol{E}_1 \\ \boldsymbol{E}_2\end{bmatrix}\boldsymbol{w}_1\left[\boldsymbol{E}_1, \boldsymbol{E}_2\right] \tag{16}$$

$$\boldsymbol{t}_{\mathrm{ep2}} = \boldsymbol{w}_2\left[\boldsymbol{E}_1, \boldsymbol{E}_2\right] \tag{17}$$

$$\boldsymbol{t} = \boldsymbol{t}_{\mathrm{ep1}} + \boldsymbol{t}_{\mathrm{ep2}} \tag{18}$$

其中权重矩阵 $\boldsymbol{w}$、$\boldsymbol{w}_1$ 和 $\boldsymbol{w}_2$ 的维度分别为 $d \times 1$、$d^{\mathrm{e}} \times 2d \times 2d$ 和 $d^{\mathrm{e}} \times 2d$，d 为Tensor-BiGRU的隐藏层维度，d^{e} 为实体对语义特征表示的输出维度．

2.5　输出层

分段注意力层和实体对语义表示层最终将输出三种固定长度的特征向量，将其拼接后的最终表示 $\boldsymbol{h}_{\mathrm{att}}$ 输入全连接层，通过softmax函数计算得到每个类别的归一化分数，其中 $\boldsymbol{w}$ 为权重矩阵，$\boldsymbol{b}$ 为偏置项．具体计算公式如下：

$$\hat{p}\left(\boldsymbol{y}\right) = \mathrm{softmax}\left(\boldsymbol{w}\boldsymbol{h}_{\mathrm{att}} + \boldsymbol{b}\right) \tag{19}$$

$$\hat{y} = \underset{y}{\mathrm{argmax}}\ \hat{p}\left(\boldsymbol{y}\right) \tag{20}$$

当前神经网络模型常用的损失函数为交叉熵：

$$\mathrm{Loss} = -\sum_{i=1}^{m} l_i \log\left(\hat{y}_i\right) + \lambda \left\| \theta \right\|^2 \tag{21}$$

其中，l_i 为训练样本的真实值，$\hat{y}_i$ 为模型对训练样本的类别预测概率，λ 为 L_2 正则化项的超参数，由此可求得模型的训练损失．传统的交叉熵损失函数建立在训练集中各类别样本均匀分布的理想状态下，然而在临床医学领域各类临床病症及其治疗案例样本的真实分布是不均衡的．以2010 i2b2/VA数据集为例，其训练集中样本数最多的类别是最少类别样本数的41倍，不同类别的训练语料间存在严重的失衡现象，这影响了模型的训练效果和泛化性．为了缓解上述问题，本文提出了权重自适应的交叉熵损失函数，根据各类别真实的训练样本分布为其赋予不同的损失惩罚权重 w_i，相关计算公式如下：

$$w_i = \begin{cases} 1 + \dfrac{n_{\mathrm{avg}} - n_i}{n}, n_i < n_{\mathrm{avg}} \\ 1 - \dfrac{n_i - n_{\mathrm{avg}}}{n}, n_i > n_{\mathrm{avg}} \end{cases} \tag{22}$$

其中，n 为训练样本总数，n_i 为第 i 个类别的样本数，n_{avg} 为训练样本类别均值．自适应的损失惩罚在一定程度上提升了小样本类别在模型训练过程中的权重，缓解

了真实训练数据中的类别不平衡问题. 另外,本文采用了基于梯度惩罚[15]的对抗训练策略,在损失函数中加入了梯度惩罚项$\frac{1}{2}\varepsilon\left\|\nabla_x \text{Loss}\right\|^2$,其中$x$为模型输入. 最终更新后的交叉熵损失函数为:

$$\text{Loss}'=-\sum_{i=1}^{m} w_i l_i \log\left(\hat{y}_i\right)+\lambda\|\theta\|^2+\frac{1}{2}\varepsilon\left\|\nabla_x \text{Loss}\right\|^2 \quad (23)$$

3 实验结果分析

本节报告了模型在2010 i2b2/VA数据集上的性能评估结果,包含与现有模型的性能对比、调参实验、消融实验、以及案例分析. 为了客观、全面地与现有研究进行比较,本文采用以下三种评估方法:准确率(P),召回率(R)和Macro-$F1$值(Macro-$F1$).

3.1 数据集

本文模型的性能评估采用2010 i2b2/VA数据集[16],它基于一系列医疗机构的电子病历数据构成,包含了患者医疗问题、治疗和测试对应的真实样本,是临床医学关系抽取领域最具代表性的公共数据集之一. 原始的2010 i2b2/VA数据集有394个训练文档和477个测试文档,共包含8种关系类型(见表1),但由于隐私原因,目前i2b2撤销了数据集中的部分文档,导致了部分类别样本缺失. 为了保证实验结果的公平性和客观性,本文遵照Sahu[7]和Li[10]等人的做法将数据集中缺失样本的类别剔除,处理后的2010 i2b2/VA数据集相关信息如表1所示,其中"*"表示被剔除后的类别.

表1 2010 i2b2/VA语料的数据统计

类别名	描述	训练集	测试集
TrCP	治疗引起医疗问题	180	335
TrAP	对医疗问题进行治疗	874	1 681
TeRP	检验显示医疗问题	981	2 035
TeCP	为调查医疗问题进行检验	165	333
PIP	医疗问题表明医疗问题	749	1 420
TrWP*	治疗恶化医疗问题	24	104
TrIP*	治疗改善医疗问题	50	151
TrNAP*	由于医疗问题不进行治疗	62	109

3.2 超参数设置

模型训练所需的各项超参数设置如表2所示. 嵌入层生成的词嵌入、位置嵌入和实体类型嵌入的维度分别为100、10和15;Tensor-BiGRU隐藏层神经单元和实体对张量表示维度分别设置为150和200;模型在训练过程中采用Adam作为优化器,将Batch Size和学习率分别设置为50和0.000 2. 另外,为了避免模型在训练时出现过拟合现象,采用Dropout和L_2两种正则化策略,参数值分别设置为0.5和0.000 1.

表2 本文模型的主要超参数

超参数	值
词嵌入维度	100
位置嵌入维度	10
实体类型嵌入维度	15
Tensor-BiGRU神经单元维度	150
张量矩阵维度	200
学习率	0.000 2
批尺寸	50
Dropout	0.5
L_2正则化系数	0.000 1

3.3 与现有模型的性能对比

为了验证本文所提方法的有效性,我们将本文模型和现有方法在2010 i2b2/VA数据集上进行了从单个类别到整体性能的细粒度比较,实验结果如表3所示. 实验结果表明,大多数深度学习方法在整体性能方面优于传统的机器学习方法,在有足够的训练标注样本的前提下,神经网络具有更好的特征学习能力. 另外,与现有的神经网络模型相比,本文模型实现了更优的关系抽取性能,Li等人[9]提出的分段注意力BiLSTM模型在2010 i2b2/VA数据集上实现了先前的SOTA性能,与之相比本文方法在整体性能上提升了0.56,而且在"TeCP"、"TrCP"、"TrAP"和"TeRP"单类别上均取得了更好的性能. 原始的CNN、RNN及其变种网络在面对句式复杂、实体对分布密集的临床医学文本无法充分地捕获潜在的有价值特征,因此在BiGRU网络的隐藏层引入张量权重矩阵后,有效地提升了模型性能,而且由于本文改进了原始的交叉熵损失函数,我们的模型在2010 i2b2/VA数据集的小样本类别上的性能提升更为明显.

3.4 调参实验

为了评价本文模型中超参数的取值对最终性能的影响,本文设置了对模型性能影响较大的几个关键超参数的调参实验,模型使用不同的超参数在2010 i2b2/VA数据集各类别中的$F1$性能表现如表4所示,其中d_{TG}表示Tensor-BiGRU中隐藏层神经单元的维度,d_{ep}表示实体对语义表示层中初始化的权重矩阵的维度,dr表示预设的Dropout比率,*N/A*表示不使用该组件. 除了上述参与实验的超参数外,模型的其余超参数根据先验知识预设,并基于验证集结果在多次实验中进行调整,具体过程不再赘述. 首先,在编码器网络维度方面,增加Tensor-BiGRU的神经单元个数能够提升模型特征提取的效果,但超过一定数量后性能表现会略有降低. 表中的1、3、6条验证了上述结论,当Tensor-BiGRU的神经单元的维度由100提高至150时,模型的性能提升到最佳,但是当维度继续提高时,模型的平均性能下降

表3　本文模型与现有方法的性能对比

Type	%	Li[9]	宁[10]	Lu[7]	Sa[6]	He[8]	Ri[4]	De[5]	Ours
TrCP	*P*	63.20	-	61.80	63.60	54.90	54.20	74.70	71.16
	R	59.80	-	40.80	43.67	42.40	56.50	32.70	56.37
	F	61.40	58.51	49.10	56.44	47.90	55.40	45.50	62.89
TrAP	*P*	82.20	-	70.60	73.49	69.70	70.70	74.80	85.35
	R	92.20	-	78.10	65.83	73.60	81.40	73.00	88.77
	F	86.90	74.37	74.20	69.23	71.60	75.70	73.90	87.02
TeRP	*P*	84.20	-	85.80	82.74	79.80	82.50	84.20	84.01
	R	91.20	-	83.10	79.88	86.50	90.60	88.00	92.22
	F	87.50	82.59	84.40	81.25	83.00	86.40	86.10	87.93
TeCP	*P*	53.70	-	68.00	63.48	70.60	59.40	85.70	54.52
	R	44.10	-	43.00	43.67	28.40	45.60	31.60	48.55
	F	48.40	59.07	52.70	50.56	40.50	51.60	46.20	51.36
PIP	*P*	66.50	-	64.00	67.32	68.90	66.40	69.10	64.96
	R	77.70	-	73.10	63.30	58.10	72.60	71.20	74.22
	F	71.60	57.94	68.30	64.92	63.10	69.40	70.10	69.28
Macro	*P*	69.96	-	70.04	70.13	68.78	66.64	77.70	72.00
	R	73.00	-	63.62	59.27	57.80	69.34	59.30	72.03
	F	71.45	66.50	66.68	64.24	62.81	67.96	67.26	72.01

*注:为方便展示,上表中对比方法的相关作者仅使用前两个字母替代.文献[20]未提供*P*和*R*结果,因此其Macro-*F*1结果取各类别*F*1值的平均数.

了0.41. 另外,我们也评估了实体对语义表示层中参与训练的权重矩阵的维度. 由表中的2、4、6条可得,当维度设置为200时,模型取得了最优性能,较维度为100和150时模型的平均性能分别提升了0.82和0.39. 更高维的权重矩阵实现了与目标实体对更充分地交互,提取和捕获了更多实体对之间的特征. 最后,我们测试了引入Dropout机制前后的模型性能比较. 由表中的5、6条可知,在保持模型的其他超参数设置一致的前提下,引入Dropout机制后模型的平均性能提升了0.54. Dropout机制可以有效地缓解模型在训练过程中的过拟合现象,在一定程度上提升模型的性能.

表4　本文模型的调参实验结果对比

d_{TG}	d_{ep}	dr	TrCP	TrAP	TeRP	TeCP	PIP	Avg
100	200	0.5	62.02	86.29	85.64	51.68	68.97	70.92
150	100	0.5	61.88	85.79	87.27	50.22	69.24	70.88
200	200	0.5	62.12	86.56	87.10	50.76	69.91	71.29
150	150	0.5	62.91	86.64	86.97	51.03	69.00	71.31
150	200	*N/A*	62.43	86.38	87.04	50.88	69.07	71.16
150	200	0.5	62.89	87.02	87.93	51.36	69.28	71.70

3.5　消融实验

为了验证本文模型中各模块的有效性,对模型进行了消融实验. 实验结果如表5所示,*w/o*表示去掉当前模块后模型在2010 i2b2/VA数据集各类别中的*F*1性能表现,TG表示引入张量权重矩阵的BiGRU网络,KaA和AP分别表示分段注意力层中的知识感知注意力和注意力池化模块,EP表示实体对的语义信息特征表示,WaL和AT分别表示输出层中的权重自适应损失和对抗训练策略. 本文在编码层设计了张量权重矩阵改进BiGRU网络,使模型的输入表示和网络隐藏层实现更充分的交互,Tensor-BiGRU的引入使模型的性能有了明显的提升. 分段注意力共包含两个子模块,其中基于知识感知的注意力模块对于性能的提升有较大的帮助,我们分析可能的原因为与基于加性注意力机制设计的分段注意力池化模块相比,考虑了外部特征的知识感知注意力模块可以更好地利用位置信息和实体类型信息. 为了缓解临床医学文本中关系实体对分布密集导致的模型性能下降问题,本文引入了实体对语义表示层,与先前的研究得出的结论一致,实体对语义表示能够使目标实体对的联系更加密切,减少冗余信息的干扰[8]. 权重自适应损失函数的引入可以提升小样本类别的训练惩罚力度,由实验结果可知,引入该损失函数后模型在“TrCP”和“TeCP”单个样本类别的关系分类性能有明显提升,这验证了本文所提方法的有效性.

表5　模型各组件的消融实验结果对比

模型	TrCP	TrAP	TeRP	TeCP	PIP	AVG
Ours	62.89	87.02	87.93	51.36	69.28	71.70
w/o TG	61.26	85.88	86.74	49.53	67.89	70.26
w/o KaA	62.19	86.02	85.76	48.77	68.16	70.18
w/o AP	62.44	86.24	86.09	50.09	68.49	70.67
w/o EP	62.08	86.29	86.71	50.66	65.91	70.33
w/o WaL	61.97	86.83	87.51	50.56	68.71	71.12
w/o AT	62.44	86.75	86.60	51.49	69.17	71.29

3.6　案例分析

为了进一步评估本文模型以及分析错误原因,本文挑选了三种能够反映2010 i2b2/VA数据集主要特点的典型例句进行分析,表6展示了本文模型在三种示例上的预测结果. 例句1为普通的单实体对-单关系类型示例,目标实体对只存在一种关系类型“TeRP”且不与示例中的其他词产生关系,本文模型可以准确地预测此类样本. 例句2为典型的多实体对-单关系类型示例,其中实体“peripheral vascular disease”与其他实体均构成关系类型“TrAP”. 虽然例句2句式冗长、结构复杂且实体对分布密集,但是本文提出的Tensor-BiGRU和分段注意力层以及引入的实体对语义信息可以有效地提升模型捕获长句特征和过滤冗余上下文语义信息的能力,因此本文模型能够对这类多实体对重叠关系进行较好的预测. 例句3虽然在句子长度和产生关系的实体对数量方面不及例句2,但是该例句为多实体对-多关系类型示例,其中实体“a cardiac catheterization”与实

体“chest pain”产生了“TeCP”关系，而与实体“an occluded right coronary artery”和“a 40-50% proximal stenosis”则产生了“TeRP”关系．本文对该示例未能准确的预测出所有实体对的关系类型，将实体“a cardiac catheterization”与实体“chest pain”的关系错误的预测为“TeRP”关系，我们分析可能的原因为实体“chest pain”和实体“an occluded right coronary artery”上下文结构类似且后面均有“and”一词，因此模型错误的预判了句式结构而导致最终的预测结果错误．另外，先前的研究也表明，属于“TeCP”关系的实体对经常被误分类为“TeRP”关系类别[8]．未来将针对关系类型重叠的复杂问题进行进一步研究．

表6 案例分析结果

	真实样本	预测结果
1	The patient had [*a MRSA nasal culture*] obtained on 06/03/05 , which revealed [*rare staphylococcus aureus*] .	The patient had [*a MRSA nasal culture*] obtained on 06/03/05 , which revealed [*rare staphylococcus aureus*] .
	{a MRSA nasal culture, TeRP, rare staphylococcus aureus}	{a MRSA nasal culture, TeRP, rare staphylococcus aureus}
2	The patient is a 64-year-old male with a long standing history of [*peripheral vascular disease*] who has had [*multiple vascular procedures*] in the past including [*a fem-fem bypass*] , [*a left fem pop*] as well as [*bilateral TMAs*] and [*a right fem pop bypass*] who presents with a nonhealing wound of his left TMA stump as well as a pretibial ulcer that is down to the bone .	The patient is a 64-year-old male with a long standing history of [*peripheral vascular disease*] who has had [*multiple vascular procedures*] in the past including [*a fem-fem bypass*] , [*a left fem pop*] as well as [*bilateral TMAs*] and [*a right fem pop bypass*] who presents with a nonhealing wound of his left TMA stump as well as a pretibial ulcer that is down to the bone .
	{multiple vascular procedures, TrAP, peripheral vascular disease}、{a fem-fem bypass, TrAP, peripheral vascular disease}、{a left fem pop, TrAP, peripheral vascular disease}、{bilateral TMAs, TrAP, peripheral vascular disease}、{a right fem pop bypass, TrAP, peripheral vascular disease}	{multiple vascular procedures, TrAP, peripheral vascular disease}、{a fem-fem bypass, TrAP, peripheral vascular disease}、{a left fem pop, TrAP, peripheral vascular disease}、{bilateral TMAs, TrAP, peripheral vascular disease}、{a right fem pop bypass, TrAP, peripheral vascular disease}
3	He has a history of [*chest pain*] and in January 1993 underwent [*a cardiac catheterization*] at Ph University Of Medical Center which revealed [*an occluded right coronary artery*] and [*a 40−50% proximal stenosis*] .	He has a history of [*chest pain*] and in January 1993 underwent [*a cardiac catheterization*] at Ph University Of Medical Center which revealed [*an occluded right coronary artery*] and [*a 40−50% proximal stenosis*] .
	{a cardiac catheterization, TeCP, chest pain}、{ a cardiac catheterization, TeRP, an occluded right coronary artery}、{ a cardiac catheterization, TeRP, a 40-50% proximal stenosis}	{a cardiac catheterization, TeRP, chest pain}、{ a cardiac catheterization, TeRP, an occluded right coronary artery}、{ a cardiac catheterization, TeRP, a 40-50% proximal stenosis}

4 结论

本文提出了一种融合了分段注意力机制的Tensor-BiGRU关系抽取模型．为了提升神经网络捕获底层特征的能力，在原始BiGRU的基础上提出了基于张量权重矩阵编码的Tensor-BiGRU网络．此外，设计了两种分段注意力机制，可以有效地提高模型捕获复杂长句特征的性能．最后，考虑到临床医学数据集中训练样本真实分布不平衡的问题，提出一种权重自适应的交叉熵损失函数，用于提升小样本类别的损失惩罚力度．实验结果表明，在保持模型灵活性和不依赖任何特征工程的前提下，本文模型在2010 i2b2/VA临床医学数据集上实现了先进的性能．在未来的工作中，我们计划专门针对关系抽取任务中的重叠实体关系的抽取等复杂问题进行研究．

参考文献

[1] 李志欣, 孙亚茹, 唐素勤, 等. 双路注意力引导图卷积网络的关系抽取[J]. 电子学报, 2021, 49(2): 315-323.
LI Z X, SUN Y R, TANG S Q, et al. Dual attention guided graph convolutional networks for relation extraction[J]. Acta Electronica Sinica, 2021, 49(2): 315-323. (in Chinese)

[2] 冯建周, 宋沙沙, 王元卓, 等. 基于改进注意力机制的实体关系抽取方法[J]. 电子学报, 2019, 47(8): 1692-1700.
FENG J Z, SONG S S, WANG Y Z, et al. Entity relation extraction based on improved attention mechanism[J]. Acta Electronica Sinica, 2019, 47(8): 1692-1700. (in Chinese)

[3] 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[J]. 自动化学报, 2014, 40(8): 1537-1562.
YANG J F, YU Q B, GUAN Y, et al. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction[J]. Acta Automatica Sinica, 2014, 40(8): 1537-1562. (in Chinese)

[4] RINK B, HARABAGIU S, ROBERTS K. Automatic extraction of relations between medical concepts in clinical texts[J]. Journal of the American Medical Informatics Association, 2011, 18(5): 594-600.

[5] DE BRUIJN B, CHERRY C, KIRITCHENKO S, et al. Machine-learned solutions for three stages of clinical information extraction: The state of the art at i2b2 2010[J]. Journal of the American Medical Informatics Association, 2011, 18 (5): 557-562.
[6] SAHU S K, ANAND A, ORUGANTY K, et al. Relation extraction from clinical texts using domain invariant convolutional neural network[EB/OL]. (2016-06-30): https://arxiv.org/abs/1606.09370.
[7] LUO Y. Recurrent neural networks for classifying relations in clinical notes[J]. Journal of Biomedical Informatics, 2017, 72: 85-95.
[8] HE B, GUAN Y, DAI R. Classifying medical relations in clinical text via convolutional neural networks[J]. Artificial Intelligence in Medicine, 2019, 93: 43-49.
[9] LI Z, YANG J S, GOU X, et al. Recurrent neural networks with segment attention and entity description for relation extraction from clinical texts[J]. Artificial Intelligence in Medicine, 2019, 97: 9-18.
[10] 宁尚明, 滕飞, 李天瑞. 基于多通道自注意力机制的电子病历实体关系抽取[J]. 计算机学报, 2020, 43(5): 916-929. NING S M, TENG F, LI T R. Multi-channel self-attention mechanism for relation extraction in clinical records [J]. Chinese Journal of Computers, 2020, 43(5): 916-929. (in Chinese)
[11] LE Q V, MIKOLOV T. Distributed representations of sentences and documents[EB/OL]. (2014-05-16): https://arxiv.org/abs/1405.4053
[12] ZENG D J, LIU K, CHEN Y B, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 1753-1762.
[13] SOCHER R, PERELYGIN A, WU J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//EMNLP 2013. Seattle: Association: for Computational Linguistics 2013: 1631-1642.
[14] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 207-212.
[15] ROSS A S, DOSHI-VELEZ F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients[EB/OL]. (2017-11-26): https://arxiv.org/abs/1711.09404.
[16] UZUNER Ö, SOUTH B R, SHEN S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text[J]. Journal of the American Medical Informatics Association, 2011, 18(5): 552-556.

作者简介

隗　昊　男,讲师,博士,1993年生于山东济南. 2017年于云南师范大学获得理学硕士学位,2021年于大连海事大学获得工学博士学位,研究方向为生物医学信息抽取、深度学习等.
E-mail: wh1005@dlmu.edu.cn

唐焕玲　女,教授,博士,硕士生导师,1970年生于山东龙口. 2004年于清华大学获得工学硕士学位,2009年于大连海事大学大学获得工学博士学位. 从事机器学习、人工智能、数据挖掘等方向的理论及应用研究.
E-mail: thl01@163.com

周　爱　女,博士研究生,1990年生于黑龙江哈尔滨. 2017年于香港理工大学获得硕士学位,现于大连海事大学攻读博士学位,研究方向为数据挖掘、机器学习等.
E-mail: zhouai9070@163.com

张益嘉　男,副教授,博士,硕士生导师,1979年生于吉林长春. 2009年于大连理工大学获得工学硕士学位,2014年于大连理工大学获得工学博士学位. 从事自然语言处理、生物医学知识挖掘等方向的理论及应用研究.
E-mail: zhangyijia@dlmu.edu.cn

陈　飞　男,讲师,博士,硕士生导师,1979年出生于辽宁本溪. 2005年于大连海事大学获得工学硕士学位,2010年于大连海事大学获得管理学博士学位. 从事人工智能、机器学习、数据挖掘等方向的理论及应用研究.
E-mail: dlmucf@163.com

鲁明羽(通讯作者)　男,教授,博士生导师. 1963年生于黑龙江鸡西. 1988年、2002年于清华大学分别获得工学硕士和工学博士学位,从事机器学习、人工智能、数据挖掘等方向的理论及应用研究.
E-mail: lumingyu@dlmu.edu.cn