基于全局上下文机制的实体关系联合抽取方法

张云涛[1] 黄 莺[1]
ZHANG Yuntao HUANG Ying

摘 要 目前，中文电子病历进行实体关系抽取存在着因医疗文本表达模糊不准确、文本结构复杂而造成医疗关系识别不准确的问题。针对这一问题，文章提出了一种基于全局上下文机制的中文电子病历实体关系联合抽取模型——GCPRel。首先，该模型通过 BERT 获取词表示。其次，借助全局上下文机制，将过去完成时和将来进行时的句子表征整合至每个单元格的句子表征内，以此更有效地捕捉文本里的上下文信息。再次，借助所获取的文本表征和词性信息，提取所有潜在的主语 - 宾语组合对。最后，通过 biaffine 模型为每个实体对分配可能的关系，得到医疗文本的三元组。采用 CHIP2020 关系抽取数据集和糖尿病数据集进行实验验证，结果显示，在 CHIP2020 关系抽取数据集上的 Precision 为 62.413%，Recall 为 60.737%，F_1 值为 61.563%；在糖尿病数据集上的 Precision、Recall 和 F_1 值分别为 83.487%、80.583% 和 82.009%，证明了该模型的三元组抽取性能优于其它基线模型。

关键词 关系抽取；联合抽取；全局上下文机制；词性信息

doi：10.3969/j.issn.1672-9528.2025.01.033

0 引言

实体关系抽取是自然语言处理中的重要任务，其在命名实体识别任务 [1] 的基础上，进一步从文本中提取出两个实体之间的关系 [2]。在医疗健康领域，这项任务尤为重要，因为医疗文本中存在着许多实体之间的复杂关联。准确识别这些关系不仅影响后续电子病历处理任务的效果和可靠性，还直接关系到病历信息的自动化分析、知识图谱的构建以及医疗决策支持系统的开发 [3]。

早期的关系抽取方法通常采用基于流水线的框架，先识别实体，再预测实体之间的关系 [4]。Socher 等人 [5] 提出了一种基于 RNN 的流水线方法，使得模型更加灵活，因为实体模型和关系模型可以使用独立的数据集，无须同时标注实体和关系的数据。然而，这些方法忽视了实体识别和关系预测之间的相关性，容易导致误差积累，进而影响后续关系抽取的性能。为了解决这些问题，越来越多的研究者开始探索联合抽取方法，以端到端的方式同时提取实体和关系，并提出了一系列新颖的方法。这些方法有效地解决了流水线方法存在的错误传播和关系依赖问题，同时有效解决了冗余实体的问题。

目前，联合抽取方法在关系抽取领域逐渐占据主导地位 [6]，根据采用的抽取方式，大致分为 3 种主要类型。

一是表格填充法。Wang 等人 [7] 利用这种方法为每种关系建立一个 $l \times l$ 的表格（l 是输入句子中的标记数），该表格中的数目通常表示两个具有特定关系实体的起始位置和结束位置（甚至是这些实体的类型）。因此，医疗文本的关系抽取任务被转化为准确且有效填充这些表格。

二是序列到序列方法。这种方法通常将三元组视为一个标记序列，并将关系抽取任务转化为生成任务，以某种顺序生成三元组。例如，Tapas 等人 [8] 在其方法中使用了编码器 - 解码器架构，使其在医疗文本中找到多个实体重叠的元组和多个标记实体的元组。Zeng 等人 [9] 提出了一种具有生成式的 transformer 对比三元组提取方法，该方法可以任意一个类的句子中联合提取相关事实。

三是基于标签的方法。这种方法通常使用二元标记序列的方式来确定实体的起始和结束位置，同时还用来确定两个实体之间的关系。Zheng 等人 [10] 提出了一种基于标签的框架，将联合抽取任务转化为一个标记问题，直接提取实体及其关系。近期，研究人员开始探索了基于单向抽取框架的标记方法，Yu 等人 [11] 首先提取所有主体，然后基于提取的主体同时提取对象和关系。Wei 等人 [12] 提出层叠式指针标注框架 CASREL，该模型通过抽取主语进而抽取宾语及其相关关系，这种单向操作会引发误差传播问题。Zheng 等人 [13] 提出了一种基于潜在关系和全局应对的联合关系三元组抽取框架 PRGC。该模型同样在处理关系重叠问题是存在误差传播的问

1. 三峡大学计算机与信息学院 湖北宜昌 443000
[基金项目] 国家重点研究发展计划资助项目“城镇安全风险评估与应急保障技术研究”（2016YFC0802500）

题。为解决这一问题，Ren 等人 [14] 提出了一种具有双向平行架构的联合抽取方法，该模型通过两个方向同时抽取主语和宾语，缓解了因为单向抽取而导致的误差传播的问题，但该模型在进行实体抽取时，忽视了实体所具有的特征，从而导致一些结构特征丢失。

然而，通用领域的关系抽取方法在医疗领域的效果有限。由于医疗文本中的实体通常为专业术语，且实体之间的关系更加复杂，简单的联合抽取方法难以充分捕捉其文本特征 [15]。

基于上述问题，本文提出了基于全局上下文机制的中文电子病历实体关系联合抽取模型——GCPRel，该模型通过两个方向同时抽取主语和宾语，缓解了因为单向抽取而导致的误差传播的问题，并且加入了全局上下文机制来获取丰富的上下文信息，通过获取到的信息以及词性信息通过两个方向帮助主语识别宾语，帮助宾语识别主语，提高模型抽取三元组的性能。

1 基于全局上下文机制的中文电子病历实体联合抽取模型

为解决中文医疗文本存在的模糊和不确定的语言表达以及复杂的语言结构进而导致模型抽取性能不佳的问题，本文提出了一种基于全局上下文机制的中文电子病历实体联合抽取模型——GCPRel。该模型通过全局上下文机制捕获更全面的文本信息，联合实体的词性信息来抽取所有可能的主语 - 宾语对，进而构建医疗三元组。本文模型由四部分组成：编码层、双向实体对抽取层（bidirectional entity pair extraction）、关系抽取层（relation extraction）和关系三元组生成层，如图 1 所示。首先，通过 BERT 层获取初始文本的原文嵌入。其次，利用全局上下文机制将整个未来和过去的句子表示整合到每个单元格的句子表示中。同时，利用文本的词性信息来辅助主语抽取任务。

1.1 编码层

1.1.1 文本特征

本文采用的预训练模型是 ERNIE-Health，该模型利用 ERNIE 先进的知识增强预训练语言模型，通过医疗知识增强技术进一步学习海量的医疗数据，精准地把握专业的医学知识。

给定输入序列 $S=\{w_1, x_2, \cdots, w_n\}$，$n$ 表示输入句子的长度，使用预训练模型生成句子初始表示 $Z=\{z_1, z_2, \cdots, z_n\}$。

1.1.2 BiLSTM 层

BiLSTM 能够对句子结构建模并保持长句子中的依赖关系，因此对于序列标注任务来说，BiLSTM 是一个非常强大的结构。在本文中，利用 BiLSTM 来增强每个单词的句子表示，以时间步长 t 为例，BiLSTM 基于 $Z=\{z_1, z_2, \cdots, z_n\}$ 生成句子表示 $\boldsymbol{H}_t$，公式为：

$$\overrightarrow{\boldsymbol{H}_t}=\overrightarrow{\mathrm{LSTM}_t}(\overrightarrow{\boldsymbol{H}_{t-1}}, z) \tag{1}$$

$$\overleftarrow{\boldsymbol{H}_t}=\overleftarrow{\mathrm{LSTM}_t}(\overleftarrow{\boldsymbol{H}_{t+1}}, z_t) \tag{2}$$

$$\boldsymbol{H}_t=\overrightarrow{\boldsymbol{H}_t} \| \overleftarrow{\boldsymbol{H}_t} \tag{3}$$

1.1.3 全局上下文机制

考虑到整个句子信息局限在 BiLSTM 的第一个和最后一个单元格中，文中使用权重 ${}^{l}\boldsymbol{H}$ 和 ${}^{l}\boldsymbol{G}$ 将其与整个句子表示 $\boldsymbol{G}=\overleftarrow{\boldsymbol{H}_1} \| \overrightarrow{\boldsymbol{H}_n}$ 合并，图 2、图 3 描述了全局上下文机制的结构。

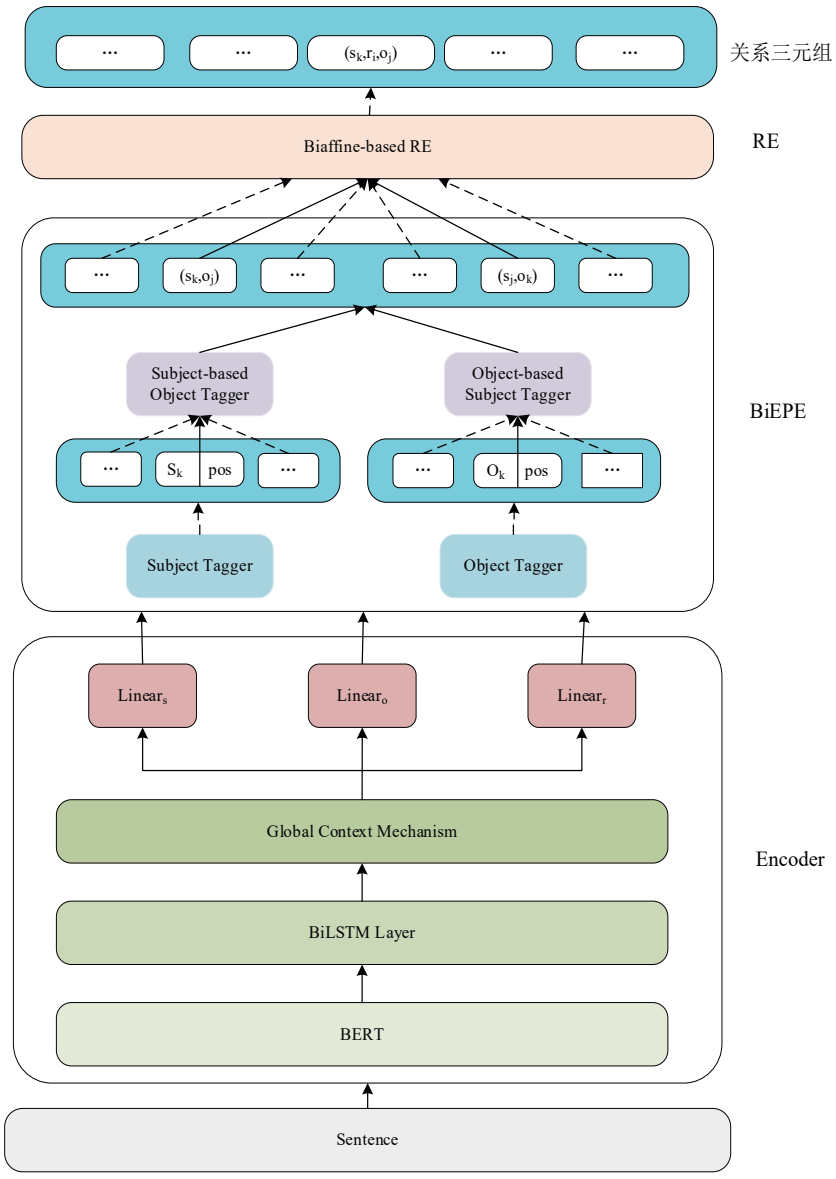

图 1 模型图

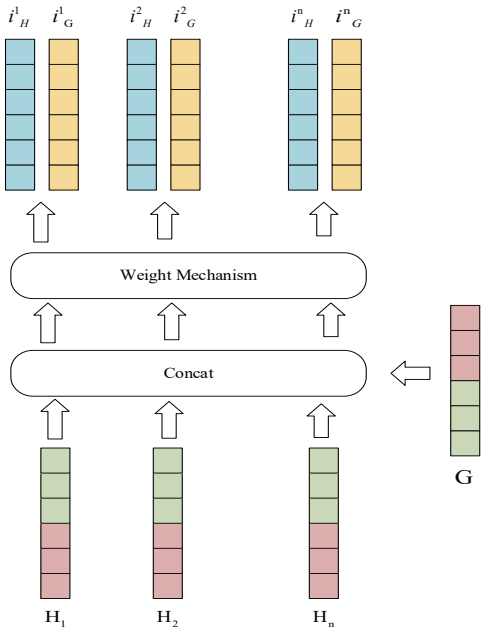

图 2 门控机制生成 i^t_H 和 i^t_G

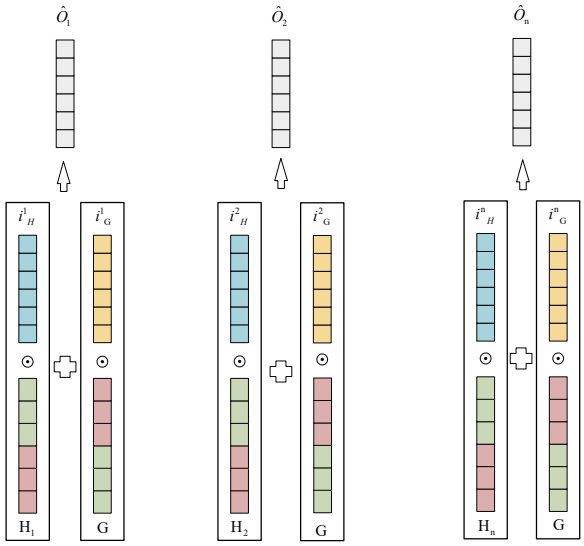

图 3 融合 i^t_H 和 i^t_G 生成全局上下文信息

给定 BiLSTM 输出 $\boldsymbol{H}=\{\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_n\}$，$\boldsymbol{H}\in\mathbf{R}^{n\times d}$；对于第 t 步，推导出 $\boldsymbol{O}_t=\boldsymbol{G}\|\boldsymbol{H}_t$，用于门控机制 [16] 生成 i^t_H 和 i^t_G，如图 2 所示。

在门控机制中，首先使用线性映射从 $\boldsymbol{O}_t$ 中选择相关的特征，过程公式为：

$$\boldsymbol{R}_H=\boldsymbol{W}_H\boldsymbol{O}_t+\boldsymbol{b}_H \tag{4}$$

$$\boldsymbol{R}_G=\boldsymbol{W}_G\boldsymbol{O}_t+\boldsymbol{b}_G \tag{5}$$

式中：$\boldsymbol{W}_H$ 和 $\boldsymbol{W}_G$ 是可训练矩阵；$\boldsymbol{R}^t_G$、$\boldsymbol{R}^t_H$ 分别代表线性变换后的全局信息 $\boldsymbol{G}$ 和当前句子表示 $\boldsymbol{H}$。

然后权重 i^t_H 和 i^t_G 分别由 sigmoid 函数计算得出：

$$i^t_H=\text{sigmoid}(\boldsymbol{R}^t_H) \tag{6}$$

$$i^t_G=\text{sigmoid}(\boldsymbol{R}^t_G) \tag{7}$$

最后，通过 $\boldsymbol{H}_t$ 和 $\boldsymbol{G}$ 融合 i^t_H 和 i^t_G 生成全局上下文信息，如图 3 所示，其过程公式为：

$$\hat{\boldsymbol{O}}_t=i^t_H\odot\boldsymbol{H}_t\|i^t_G\odot\boldsymbol{G} \tag{8}$$

式中：⊙表示元素乘积。

利用获取的全局上下文信息，生成 3 种不同的标记序列，作为主语、宾语和关系的上下文特征表示，它们分别为 $\boldsymbol{O}^i_s$、$\boldsymbol{O}^i_o$、$\boldsymbol{O}^i_r$，其公式为：

$$\boldsymbol{O}^i_s=\boldsymbol{W}_s\boldsymbol{O}^i+\boldsymbol{b}_s \tag{9}$$

$$\boldsymbol{O}^i_o=\boldsymbol{W}_o\boldsymbol{O}^i+\boldsymbol{b}_o \tag{10}$$

$$\boldsymbol{O}^i_r=\boldsymbol{W}_r\boldsymbol{O}^i+\boldsymbol{b}_r \tag{11}$$

式中：$\boldsymbol{W}_{(\cdot)}$ 是一个可训练矩阵；$\boldsymbol{b}_{(\cdot)}$ 是偏置向量。

此外，考虑到三元组中的主语和宾语是高度相关的，一个实体的特征将会有助于对于另一个实体的提取。因此，将宾语表示序列的 CLS 向量（表示为 $\boldsymbol{O}^{\text{cls}}{}_o$）添加到 $\boldsymbol{O}^i_s$ 中，以增强主语的上下文特征。在宾语部分也进行相同的操作，其公式为：

$$\boldsymbol{O}^i_s=\boldsymbol{O}^i_s+\boldsymbol{O}^{\text{cls}}{}_o \tag{12}$$

$$\boldsymbol{O}^i_o=\boldsymbol{O}^i_o+\boldsymbol{O}^{\text{cls}}{}_s \tag{13}$$

1.2 双向实体对抽取（BiEPE）

BiEPE 具有双向框架，从两个方向提取主语 - 宾语对。一是 s2o 方向，首先提取主语，然后根据主语提取宾语；二是 o2s 方向，一个反向的提取，首先提取宾语，然后根据宾语提取主语。这两个方向的提取共享编码器组件。这两个方向的内部结构相似，通过 s2o 方向，能够推出 o2s 方向。

Subject Tagger 是一个基于二进制标记的模块，旨在从输入句子中提取所有的主语。对于输入句子中的每个标记，分别分配两个概率，用于表示它作为主语起始标记和结束标记的可能性，概率公式为：

$$\boldsymbol{p}^{i,\text{start}}{}_s=\sigma(\boldsymbol{W}^{\text{start}}{}_s\boldsymbol{O}^i_s+\boldsymbol{b}^{\text{start}}{}_s) \tag{14}$$

$$\boldsymbol{p}^{i,\text{end}}{}_s=\sigma(\boldsymbol{W}^{\text{end}}{}_s\boldsymbol{O}^i_s+\boldsymbol{b}^{\text{end}}{}_s) \tag{15}$$

式中：$\boldsymbol{p}^{i,\text{start}}{}_s$和$\boldsymbol{p}^{i,\text{end}}{}_s$分别表示第 i 个标记作为主语起始和结束的概率；$\boldsymbol{W}^{(\cdot)}{}_s$ 是一个可训练矩阵；$\boldsymbol{b}^{(\cdot)}{}_s$ 是偏置向量。本文所有公式中，σ 表示 sigmoid 激活函数。

文中采用简单的 1/0 标记方案，即如果一个标记的概率超过一定阈值，则将其分配为 1 标记，否则为 0 标记。

Subject-based Object Tagger 是被用于在提取到的主语的条件下提取所有宾语。在这一步，将获取到的主语词性与主语的向量表示结合帮助模型更准确地识别和提取与选定主语相关的宾语。具体而言，文中将词性信息作为额外的特征输入到模型中，在迭代标记结构中，每个标记的两个概率（作

为与选定主语相关宾语的起始和结束标记的可能性）将受到词性信息的影响。因此，模型在决定每个标记是否属于宾语的起始或结束时，可以更充分地考虑到标记本身的语法特征和上下文信息，从而提高了宾语的识别精度和准确性，概率公式为：

$$\boldsymbol{v}^{\mathrm{s\_k}}{}_s=\mathrm{maxpool}(\boldsymbol{O}^{\mathrm{s\_k\_start}}{}_s,...,\boldsymbol{O}^{\mathrm{s\_k\_end}}{}_s) \tag{16}$$

$$\boldsymbol{p}^{\mathrm{s\_k}}{}_s=\mathrm{padding}(\boldsymbol{p}^{\mathrm{s\_k\_start}}{}_s,...,\boldsymbol{p}^{\mathrm{s\_k\_end}}{}_s) \tag{17}$$

$$\boldsymbol{v}^{\mathrm{s\_k}}{}_s=\sum_{i=0}^{k}\boldsymbol{v}^{\mathrm{s\_i}}{}_s+\boldsymbol{p}^{\mathrm{s\_i}}{}_s \tag{18}$$

$$\boldsymbol{p}^{i,\mathrm{start}}{}_o=\sigma(\boldsymbol{W}^{\mathrm{start}}{}_o(\boldsymbol{O}^{i}{}_o\times\boldsymbol{v}^{\mathrm{s\_k}}{}_s)+\boldsymbol{b}^{\mathrm{start}}{}_o) \tag{19}$$

$$\boldsymbol{p}^{i,\mathrm{end}}{}_o=\sigma(\boldsymbol{W}^{\mathrm{end}}{}_o(\boldsymbol{O}^{i}{}_o\times\boldsymbol{v}^{\mathrm{s\_k}}{}_s)+\boldsymbol{b}^{\mathrm{end}}{}_o) \tag{20}$$

式中：$\boldsymbol{O}^{\mathrm{s_k_start}}{}_s,...,\boldsymbol{O}^{\mathrm{s_k_end}}{}_s$是第 k 个主语中标记的向量表示，因此$\boldsymbol{v}^{\mathrm{s_k}}{}_s$可以被视为第 k 个主语的表示；$\boldsymbol{p}^{\mathrm{s_k_start}}{}_s,...,\boldsymbol{p}^{\mathrm{s_k_end}}{}_s$是第 k 个主语中标记的词性向量表示，通过 padding 操作，使得词性信息能够与主语表示对齐，从而实现词性信息辅助模型识别与选定主语相关的宾语；maxpool(·) 表示最大池化操作；$\boldsymbol{p}^{i,\mathrm{start}}{}_o$和$\boldsymbol{p}^{i,\mathrm{end}}{}_o$分别表示与第 k 个主语相关的宾语的起始和结束标记的概率；* 表示对应元素相乘；$\boldsymbol{W}^{(\cdot)}{}_o$ 一个可训练矩阵；$\boldsymbol{b}^{(\cdot)}{}_o$ 是一个偏置向量。

由于上述两个任务中的所有的提取模块都以多任务学习的方式工作。因此，每个方向的提取模块都有自己的损失函数，将上述两个标签器模块中的损失分别表示为 L_{s1} 和 L_{o1}，使用基于二元交叉熵损失函数来定义它们，其公式为：

$$\mathrm{ce}(\boldsymbol{p},\boldsymbol{t})=-[\boldsymbol{t}\log\boldsymbol{p}+(1-\boldsymbol{t})\log(1-\boldsymbol{p})] \tag{21}$$

$$L_{s1}=\frac{i}{2\times l}\sum_{m\in\{\mathrm{start,end}\}}\sum_{i=1}^{l}\mathrm{ce}(\boldsymbol{p}^{i,m}{}_s,\boldsymbol{t}^{i,m}{}_s) \tag{22}$$

$$L_{o1}=\frac{i}{2\times l}\sum_{m\in\{\mathrm{start,end}\}}\sum_{i=1}^{l}\mathrm{ce}(\boldsymbol{p}^{i,m}{}_o,\boldsymbol{t}^{i,m}{}_o) \tag{23}$$

式中：$\mathrm{ce}(\boldsymbol{p},\boldsymbol{t})$ 为二元交叉熵损失函数，$p\in(0,1)$ 是预测概率；t 是真实标签；l 为输入句子的标签数目。

同样，在 o2s 方向有两个标记损失值，分别记为 L_{s2} 和 L_{o2}，计算方式与公式（22）（23）相似。

1.3 关系抽取（RE）

由于 BiEPE 模块输出许多的主语 - 宾语对，导致存在许多噪声对。这对模型的精度是有害的。因此，RE 应该具有强大的分类能力。这里使用一个 biaffine 模型作为 RE 模块。它为每个关系维护一个参数矩阵，而每个实体对将会使用对应关系的矩阵进行计算，以确定它是否具有相应的关系。具体来说，对于一个实体对 (s_k, o_j)，首先获得其两个实体的表示向量$\boldsymbol{v}^{\mathrm{s_k}}{}_r$和$\boldsymbol{v}^{\mathrm{o_k}}{}_r$。然后计算 (s_k, o_j) 具有第 i 个关系的可能性，表示为$\boldsymbol{p}^{i}{}_r$。过程公式为：

$$\boldsymbol{v}^{\mathrm{s\_k}}{}_r=\mathrm{maxpool}(\boldsymbol{O}^{\mathrm{s\_k\_start}}{}_r,...,\boldsymbol{O}^{\mathrm{s\_k\_end}}{}_r) \tag{24}$$

$$\boldsymbol{v}^{\mathrm{o\_j}}{}_r=\mathrm{maxpool}(\boldsymbol{O}^{\mathrm{o\_j\_start}}{}_r,...,\boldsymbol{O}^{\mathrm{o\_j\_end}}{}_r) \tag{25}$$

$$\boldsymbol{p}^{i}{}_r=\sigma\left(\begin{bmatrix}\boldsymbol{v}^{\mathrm{s\_k}}{}_r\\1\end{bmatrix}^{\mathrm{T}}\boldsymbol{W}^{i}{}_r\begin{bmatrix}\boldsymbol{v}^{\mathrm{o\_j}}{}_r\\1\end{bmatrix}\right) \tag{26}$$

式中：$\boldsymbol{W}^{i}{}_r\in\mathbf{R}^{(d_h+1)\times(d_h+1)}$是第 i 个关系的参数矩阵。

选择 biaffine 模型因其具有两大优点。一是为每个关系建立并维护一个矩阵，可以准确地建模关系特征。二是其概率计算机制使得它能够准确地挖掘主语和宾语之间的相互关系。为了训练 RE 组件，定义了一个基于交叉熵的损失函数：

$$L_r=\frac{1}{|R|}\sum_{i=1}^{|R|}\mathrm{ce}(\boldsymbol{p}^{i}{}_r,\boldsymbol{t}^{i}{}_r) \tag{27}$$

式中：R 是预定义的关系集；$|R|$ 是所有关系的数目。

1.4 损失函数

模型包括 5 个提取模块和 1 个编码器模块。除了一个特定任务使用原始句子作为输入外，其他任务都采用了 teacher forcing 模式进行训练，即通过提供正确样本来指导模型生成预测 [17]。为了解决暴露偏差问题，随机生成的负样本被合并到正样本中，以模拟真实推断阶段的情景，从而训练出更稳健的模型。模型的损失函数为：

$$L=L_{s1}+L_{o1}+L_{s2}+L_{o2}+L_r \tag{28}$$

考虑到编码器模块接收到来自不同提取模块的反向传播梯度，其收敛速度可能会与其他提取模块不同步。导致收敛速度不一致的问题，即如果使用统一学习率，一些模块可能会过度训练，而另一些模块则训练不足。因此，文中采用了一个共享感知学习机制，为不同模块分配不同的学习率。即一个模块被共享的任务越多，则为其分配更小的学习率越小。学习机制使用公式分配学习率。

$$\xi_i=\begin{cases}\xi, & k_i=1\\ \dfrac{(1+\delta)}{f(k_i)}\times\xi, & k_i>1\end{cases} \tag{29}$$

式中：ξ 是基础学习率；ξ_i 是第 i 个模块的学习率；k_i 是第 i 个模块被共享的任务数量；$\delta\in[0,1]$ 是调节因子，用于调整学习率；$f(\cdot)$ 是映射函数，将输入的 k_i 转换为合理的实数值（通常大于 1），以确定学习率的主要幅度。

2 实验

2.1 数据集

为验证本文提出的方法的有效性，分别在 CHIP2020 中文医疗文本实体关系抽取数据集和糖尿病数据集上进行了实验。

CHIP2020 数据集有近 7.5 万三元组数据，2.8 万疾病语句以及 53 种预定义关系类别，训练语料均来自专业医生编写的教材 [18]。部分关系三元组 schemas 如表 1 所示。

表 1 CHIP2020 数据集部分关系三元组 schemas

头实体类型	关系类型	尾实体类型
疾病	预防	其他
	阶段	其他
	就诊科室	其他
	辅助治疗	其他治疗
流行病学	同义词 (流行病学 / 流行并学)	流行病学
社会学	同义词 (社会学 / 社会学)	社会学
部位	同义词 (部位 / 部位)	部位
手术治疗	同义词 (手术治疗 / 手术治疗)	手术治疗

糖尿病数据集来自某市疾控中心，经过脱敏处理。部分关系三元组 schemas 如表 2 所示。

表 2 糖尿病数据集部分关系三元组 schemas

头实体类型	关系类型	尾实体类型
并发症	检查症状 - 并发症	检查症状
	自我监测症状 - 并发症	自我监测症状
	手术 - 并发症	手术
疾病	手术 - 疾病	手术
	药物 - 疾病	药物
	检查症状 - 疾病	检查症状

CHIP2020 数据集和糖尿病数据集的基本信息如表 3 所示。

表 3 数据集基本信息

数据集	训练集	验证集	实体类型	关系类别
CHIP2020 关系抽取数据集	14 339	3585	11	53
糖尿病数据集	926	204	8	13

2.2 评价指标

本文评价指标采用准确率（Precision）、召回率（Recall）和 F_1 值。计算公式为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{30}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{31}$$

$$F_1 = 2 \times \text{Precision} \times \frac{\text{Recall}}{\text{Precision+Recall}} \tag{32}$$

式中：TP（true positive）为正样本判为正的数量；FP（false positive）为正样本判为负的数量；FN（false negative）为负样本判为正的数量。

2.3 实验环境与参数设置

为找到最适合模型的学习率，在糖尿病数据集上使用了不同的学习率进行训练和评估，最终选择了模型表现最佳的 1e-5 学习率作为本文训练的学习率。实验超参数如表 4 所示。

表 4 实验超参数设置

数据集	CHIP2020 关系抽取数据集	糖尿病数据集
批处理大小	2	4
学习率	1e-5	1e-5
句子长度	256	256
最大训练次数	100	200
BiLSTM 隐藏层维度	768	768
模型优化器	Adam	Adam

2.4 实验结果分析

在 CHIP2020 关系抽取数据集和糖尿病数据集上进行对比实验，以验证 GCPRel 模型的性能，表 5 和表 6 分别列出了 CHIP2020 数据集和糖尿病数据集的基线模型，包括基于潜在关系和全局对应的联合关系三重抽取模型 PRGC；将实体关系联合抽取任务表述为集合预测问题的 SPN 模型 [19]；利用关系与标记之间的全局关联，通过迭代的方式逐步优化每个关系的表格特征进行关系三元组抽取的模型 GRTE[20]；使用级联二进制方法，先识别主语，再识别特定关系下的宾语，最终抽取三元组的模型 CASREL；采用双向架构平行抽取主语 - 宾语和宾语 - 主语对，通过 biaffine 识别关系来抽取三元组的模型 BiRTE；利用对抗训练生成对抗样本，并融合主语、宾语和关系特征辅助三元组抽取的模型 AMFRel[22]。表 5 和表 6 结果表明，GCPRel 模型在 chip2020 关系抽取数据集中取得了 61.563% 的 F_1 值，超越了该数据集提供的基线模型，与基于双向架构的关系抽取模型 BiRTE 相比，本文模型在准确率、召回率和 F_1 值 3 个指标都得到了明显的提升。GCPRel 模型在糖尿病数据集中取得了 82.009% 的 F_1 值，超越了该数据集提供的基线模型，与基于双向架构的关系抽取模型 BiRTE 相比，本文模型在准确率、召回率和 F_1 值 3 个指标都得到了明显的提升。这表明，在模型中融入全局上下文机制和词性特征对提升三元组的抽取性能是有效的。

表 5 CHIP2020 数据集上各模型结果

单位：%

	模型	Precision	Recall	F_1
Data from CHIP2020 数据集	BERT			54.0
	RoETa			56.4
	MacBERT			53.2
	PCL			49.1
Data from this work	PRGC	52.626	49.098	50.801
	SPN	56.330	49.977	52.964
	GRTE	60.563	57.213	58.840
	$\text{CASREL}_{\text{ERNIE}}$	63.401	54.763	58.766
	$\text{AMFRel}_{\text{ERNIE}}$	63.922	57.279	60.418
	BiRTE	60.049	60.416	60.232
	GCPRel	62.413	60.737	**61.563**

表 6 糖尿病数据集上各模型结果

单位：%

	模型	Precision	Recall	F_1
Data from this work	PRGC	73.169	64.648	68.645
	SPN	77.200	63.175	69.487
	GRTE	81.687	64.975	72.084
	$CASREL_{ERNIE}$	79.452	66.448	72.371
	$AMFRel_{ERNIE}$	83.914	67.021	74.145
	BiRTE	79.516	79.774	79.645
	GCPRel	83.487	80.583	**82.009**

2.5 消融实验

为验证所提出模型中关键组件的有效性，在糖尿病数据集上进行了消融实验，实验结果如表 7 所示。

表 7 消融实验结果

单位：%

模型	Precision	Recall	F_1
本文模型	83.487	80.583	82.009
本文模型 - 全局上下文机制	80.998	81.392	81.195
本文模型 - 词性特征	82.128	79.935	81.017

首先，移除全局上下文机制后，模型在精确度方面下降了 2.489%，导致整体性能下降了 0.814%。实验结果表明移除全局上下文机制会降低模型的预测性能，这是因为全局上下文机制提供了更多的文本语境信息，包括句子之间的关联和逻辑连贯性，移除它会使模型对文本的理解变得更局限，从而降低三元组抽取的效果。其次，移除词性特征后，模型在精确度方面下降了 1.359%，这也导致了整体性能下降了 0.992%，这是因为词性信息能够帮助模型更好地理解实体之间的关系，通过识别实体的词性，模型可以更有效地确定不同实体间的语法关系从而提高关系预测的准确性，而移除词性信息后，模型无法深入理解语义，从而导致三元组抽取效果不佳。

3 结语

针对医疗文本关系抽取时，存在的不确定的语言表达和复杂的语言结构进而导致模型抽取性能不佳的问题，提出了一种基于全局上下文机制的中文电子病历实体关系联合抽取模型 GCPRel。该模型通过全局上下文机制获取更详细的上下文信息，同时双向框架分别提取主语 - 宾语对和宾语 - 主语对，最终通过 Biaffine 模块提取出医疗文本中的三元组。此外，模型还利用词性信息辅助主语识别宾语，提高三元组的识别性能。实验结果表明，该模型在 F_1 值上的表现优于其他模型，证明了 GCPRel 模型能够有效识别中文医疗文本中的复杂关系。在未来的研究中，将尝试对最后的关系判断部分进行优化，以增强关系抽取效果。

参考文献：

[1] 周佳伦，李琳宇，马洪彬，等 .MRC-PBM: 一种中文电子病历嵌套命名实体识别方法 [J]. 国外电子测量技术，2024, 43(1): 159-165.

[2] LANDOLSI M Y, HLAOUA L, ROMDHANE L B. Extracting and structuring information from the electronic medical text: state of the art and trendy directions[J].Multimedia tools and applications, 2023, 83(7): 21229-21280.

[3] PAN S R, LUO L H, WANG Y F, et al. Unifying large language models and knowledge graphs: a roadmap[J].IEEE transactions on knowledge and data engineering. 2024, 36(7): 3580-3599.

[4] CHEN Y S, ROTH D. Exploiting syntactico-semantic structures for relation extraction[C]//49th Annual meeting of the Association for Computational Linguistics 2011. Stroudsburg, PA: ACL, 2011:551-560.

[5] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces[C]// Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012, vol.2. Stroudsburg, PA: ACL, 2012:1201-1211.

[6] ZHAO X Y, DENG Y, YANG M, et al. A comprehensive survey on deep learning for relation extraction: recent advances and new frontiers[J].ACM computing surveys, 2024,56(11):1-39.

[7] WANG Y C, YU B W, ZHANG Y Y, et al. TPLinker: single-stage joint extraction of entities and relations through token pair linking[C]//Proceedings of the 28th International Conference on Computational Linguistics.Barcelona：ICCL, 2020: 1572-1582.

[8] NAYAK T, NG H T. Effective modeling of encoder-decoder architecture for joint entity and relation extraction[DB/OL].(2019-11-22)[2024-07-06]https://doi.org/10.48550/arXiv.1911.09886.

[9] ZENG X R, ZENG D J, HE S Z, et al. Extracting relational facts by an end-to-end neural model with copy mechanism[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2018: 506-514.

[10] ZHENG S C, WANG F, BAO H Y, et al. Joint extraction of entities and relations based on a novel tagging scheme[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA:ACL, 2017:1227-1236.

[11] YU B W, ZHANG Z Y, SHU X B, et al. Joint extraction of entities and relations based on a novel decomposition

strategy[DB/OL].(2019-09-10)[2024-07-25].https://doi.org/10.48550/arXiv.1909.04273.

[12] WEI Z P, SU J L, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[DB/OL].(2019-09-07)[2024-07-19].https://doi.org/10.48550/arXiv.1909.03227.

[13] ZHENG H Y, WEN R, CHEN X, et al. PRGC: potential relation and global correspondence based joint relational triple extraction[DB/OL].(2021-06-18)[2024-06-19].https://doi.org/10.48550/arXiv.2106.09895.

[14] REN F L, ZHANG L H, ZHAO X F, et al. A simple but effective bidirectional framework for relational triple extraction[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. NewYork:ACM, 2022: 824-832.

[15] ZHANG X L, ZHANG T, YAN R. Chinese medical intent recognition based on multi-feature fusion[J]. Neural information processing, 2023, 1962(11): 504-515.

[16] SUN C A, LI T J, LIU T L, et al. A joint model based on interactive gate mechanism for spoken language understanding[J]. Applied intelligence, 2021, 52(8): 6057-6064.

[17] LI S Y, WANG T T, LI G Y, et al. Short-term ship roll motion prediction using the encoder–decoder Bi-LSTM with teacher forcing[J]. Ocean engineering, 2024,295(3): 116917.

[18] 甘子发，昝红英，关同峰，等 .CHIP 2020 评测任务 2 概述：中文医学文本实体关系抽取 [J]. 中文信息学报，2022, 36(6): 101-108.

[19] SUI D B, ZENG X R, CHEN Y B, et al. Joint entity and relation extraction with set prediction networks[J]. IEEE transactions on neuralnetworks and learning systems, 2023, 35(9): 12784-12795.

[20] REN F L, ZHANG L H, YIN S J, et al. A novel global feature-oriented relational triple extraction model based on table filling[DB/OL]. (2021-09-14)[2024-05-25].https://doi.org/10.48550/arXiv.2109.06705.

[21] 余肖生，李琳宇，周佳伦，等 . AMFRel: 一种中文电子病历实体关系联合抽取方法 [J]. 重庆理工大学学报 (自然科学), 2024,38(2):189-197.

【作者简介】

张云涛（1999—），通信作者（email: 964447582@qq.com），男，湖北武汉人，硕士研究生，研究方向：大数据分析技术研究。

黄莺（2000—），女，湖北武汉人，硕士研究生，研究方向：大数据分析技术研究，email: 156376342@qq.com。

（收稿日期：2024-09-02）

（上接第 139 页）

3 结语

本文提出了一种分层嵌入大模型架构，该架构以大模型技术为基础，提供了一个高效、智能的解决方案，以应对当前多模态健康医疗数据治理中所面临的复杂挑战。通过编排调度引擎将专能组件进行编排组合，形成数据治理各环节中的智能体，实现了对多模态数据的智能化处理，并成功应用于电子病历解析和医学影像分析。事实证明，该框架可为数据治理中的复杂任务提供更加精确、智能、高效的解决方案。

未来，随着大模型技术的进一步发展和医疗数据的不断积累，基于分层嵌入大模型的智能治理架构有望在更多的临床场景中发挥作用，为精准医疗、个性化健康管理等领域提供重要的技术支撑。

参考文献：

[1] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[DB/OL].(2021-06-03)[2024-02-18].https://doi.org/10.48550/arXiv.2010.11929.

[2] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[DB/OL].(2019-05-24)[2024-02-06].https://doi.org/10.48550/arXiv.1810.04805.

[3] 阮彤，邱加辉，张知行，等 . 医疗数据治理 : 构建高质量医疗大数据智能分析数据基础 [J]. 大数据，2019(1):12-24.

[4] 刘志红 . 人工智能大模型的隐私保护与数据安全技术研究 [J]. 软件，2024, 45(2): 143-145+151.

【作者简介】

马良（1989—），男，山东泰安人，硕士研究生，高级工程师，研究方向：大数据、人工智能、数据治理。

马俊朋（1979—），男，山东泰安人，本科，工程师，研究方向：大数据、人工智能。

李向阳（1981—），男，山东济南人，本科，高级工程师，研究方向：大数据、物联网、元宇宙、人工智能。

谢超（1983—），男，山东枣庄人，本科，工程师，研究方向：大数据、人工智能、数据治理。

刘超（1993—），男，山东临沂人，博士，工程师，研究方向：生物医学图像处理、生成式大模型应用研究。

（收稿日期：2024-09-10）