计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):279−287 [doi: 10.15888/j.cnki.csa.008747]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于位置特征和句法依存树的可度量数量信息抽取模型①

聂文杰[1], 莫　迪[2], 黄邦锐[2], 刘　海[1], 郝天永[1]

[1](华南师范大学 计算机学院, 广州 510631)
[2](华南师范大学 人工智能学院, 佛山 528225)
通信作者: 郝天永, E-mail: haoty@m.scnu.edu.cn

摘　要: 随着医疗信息化水平的不断提高, 电子病历得到了越来越广泛的应用, 其中的非结构化文本包含大量蕴含患者病况信息的可度量数量信息, 由于实体与数量信息表述的复杂性, 从非结构化电子病历文档中精准抽取可度量数量信息是一个重要的挑战. 本文基于双向门控循环单元提出了结合相对位置特征与注意力机制的 RPA-GRU 模型, 通过将相对位置特征融入注意力机制更新双向门控循环单元输出, 识别实体与数量信息. 并基于重构句法依存树的图注意力网络学习图级表示提出 GATM 模型, 实现实体与数量信息的关联. 实验基于 1 359 份三甲医院烧伤科电子病历数据, 结果表明 RPA-GRU 模型与 GATM 模型在可度量数量信息识别和关联上分别获得 97.58% 与 97.86% 的 *F*1 值, 比表现最好的基线模型分别高出 2.17% 与 1.74%, 验证了所提出模型的有效性.

关键词: 可度量数量信息; 电子病历; 相对位置特征; 句法依存树; 图注意力网络; 信息抽取

引用格式: 聂文杰,莫迪,黄邦锐,刘海,郝天永.基于位置特征和句法依存树的可度量数量信息抽取模型.计算机系统应用,2022,31(10):279–287. http://www.c-s-a.org.cn/1003-3254/8747.html

Extraction Model of Measurable Quantitative Information Based on Position Feature and Dependency Tree

NIE Wen-Jie[1], MO Di[2], HUANG Bang-Rui[2], LIU Hai[1], HAO Tian-Yong[1]

[1](Shool of Computer Science, South China Normal University, Guangzhou 510631, China)
[2](School of Artificial Intelligence, South China Normal University, Foshan 528225, China)

Abstract: As medical informatization is constantly improving, electronic medical records have been more and more widely used, of which the unstructured text contains massive measurable quantitative information including patient clinical conditions. Due to the complexity of entities and quantitative information, it is a challenge to accurately extract measurable quantitative information. In this study, we propose the RPA-GRU model combining the relative position feature and attention mechanism based on a bi-directional gated recurrent unit. It incorporates the relative position feature into the attention mechanism to identify entities and quantity information. Meanwhile, the GATM model is proposed according to the reconstructed dependency tree-based graph attention network to learn graph-level representation, thus achieving the association between entities and quantity information. The experiment is based on 1 359 electronic medical records from the burn injury department of a three-A hospital. The results show that the *F*1 values of RPA-GRU model and GATM model are 97.58% and 97.86% respectively in terms of identification and association of measurable quantitative information, up by 2.17% and 1.74% compared with the best-performing baseline model. In this way, the effectiveness of the proposed models is validated.

Key words: measurable quantitative information; electronic medical records; relative position feature; dependency tree; graph attention networks; information extraction

① 基金项目: 广东自然科学基金 (2021A1515011339)
收稿时间: 2022-01-17; 修改时间: 2022-02-17; 采用时间: 2022-03-03; csa 在线出版时间: 2022-06-24

随着电子病历的快速普及与发展, 从电子病历中抽取所需关键信息逐渐成为医学信息学领域研究者关注的热点问题, 目前许多研究者关注于从非结构化电子病历文本中抽取医学概念[1]、医学属性值[2]、时间表达式[3]、药物不良反应事件[4] 与药物间相互作用[5]. 对电子病历中的可度量数量信息的抽取却较为匮乏. 可度量数量信息广泛存在于各类非结构化文本中[6], 例如在临床试验钠排标准文本中的占比超过 40%[7]. 低精度的可度量数量信息抽取会导致药物剂量分析与临床试验资格标准认定等研究的瓶颈[6].

可度量数量信息作为一种量化数据, 由实体与相关数量属性组成[8]. 以语句“心率达 120 次/分钟”为例, 其中“心率”为实体, “120”为数值, “次/分钟”为单位, 数值与单位的组合“120 次/分钟”为数量. 图 1 显示了非结构化电子病历文本包含的可度量数量信息, 其中下划线表示实体, 粗体表示数值, 斜体表示单位, 其中实体与数值、单位之间的相对位置并不固定, 以“体温 36.0 摄氏度”与“3 600 mL 血浆”为例, 其中“体温 36.0 摄氏度”中的实体在数值与单位之前, 而“3 600 mL 血浆”中的实体在数值与单位之后. 另外如实体“5% 葡萄糖注射液”所示, 部分数值信息为实体的一部分, 而非单独的数值. 现有信息抽取技术尚未对可度量数量信息中的位置信息进行深入的研究, 并且难以区分单独的数值与作为实体一部分的数值.

术中详细情况见手术记录, 术中患者失血少量, 术中输乳酸钠林格氏液 **1 000** *mL*, 生理盐水 **100** *mL*, 血浆**1 200** *mL*, 患者生命体征保持稳定。术后患者体温 **36.0**摄氏度, 心率达 **120** 次/分钟。嘱咐术后密切监察切口渗血情况。继续输 **3 600** *mL* 血浆、**3 500** *mL* 乳酸钠林格氏液、5% 葡萄糖注射液 **2 000** *mL* 抗休克治疗, 注意监测生命体征变化。

图 1 非结构化电子病历文本中包含的可度量数量信息

现有可度量数量信息抽取相关研究主要利用基于规则与传统机器学习模型的方法, 然而基于规则的方法需要花费大量时间与精力设计规则, 且泛用性往往较弱, 无法很好地迁移至其他语料或领域. 而传统机器学习模型需要做大量的特征工程, 所生成的特征质量很大程度地影响着模型的最终性能. 因此可以自动抽取特征的深度学习模型引起了研究者的关注, 循环神经网络 (recurrent neural network, RNN) 被引入用来抽取信息, 同时为了进一步提升模型性能, 诸如位置特征等外部特征被融入到深度学习模型当中. 然而无论是 Vaswani 等[9] 根据 sin 函数与 cos 函数生成的位置编码还是 Wang 等[10] 介绍的位置向量, 都没有对所需信息与无关信息进行特殊处理. 此外当前大多研究将整个序列作为模型的输入, 而 Zhang 等[11] 已经证明对原输入序列进行适当删减有助于提升模型性能.

本文首先通过相对位置特征来区分实体与数量信息与非实体与非数量信息, 并将其融入注意力 (attention) 机制中, 对通过双向门控循环单元 (bi-direction gated recurrent unit, BiGRU) 获得的上下文特征进行更新, 以此识别实体与数量信息. 并通过将输入语句转换为句法依存树的同时进行重构, 在充分提取输入语句语义信息的同时排除无关信息的干扰, 并结合图注意力网络 (graph attention networks, GAT) 进一步抽取特征, 对实体与数量进行正确关联, 实现可度量数量信息关联, 最终完成可度量数量信息的抽取. 综上所述, 本文的主要贡献如下:

(1) 通过将相对位置特征与注意力机制融合, 提出新的 RPA-GRU (relative position attention-BiGRU) 模型, 识别实体与数量信息.

(2) 通过对输入语句生成的句法依存树重构, 提出新的 GATM (graph attention networks for measurable quantitative information) 模型, 关联可度量数量信息.

(3) 实验结果表明所提出的 RPA-GRU 与 GATM 模型相比基线模型获得了最佳性能, 验证了其有效性.

1 相关工作

对于可度量数量信息抽取的相关研究, 早期为基于规则的方法, 如肖洪等[12] 通过对量词进行总结得到 125 种模式, 在利用有限自动机抽取量词的同时构建正则表达式与模板从年鉴文本当中抽取数值知识元. Turchin 等[13] 利用正则表达式从临床笔记当中抽取血压值, 并通过领域知识校验抽取结果. Hao 等[7] 引入领域知识与 UMLS 元词典等外部知识设计启发式规则从 1 型糖尿病数据集与 2 型糖尿病数据集中抽取可度量数量信息. Liu 等[8] 对医学文本当中的关键语义角色进行标记, 自动学习模式抽取可度量数量信息以减少人工. 随着传统机器学习的发展, 如条件随机场 (conditional random field, CRF) 被引入, 或单独使用或与规则进行结合. 张桂平等[14] 在构建模板的基础上利用 CRF 对模板进行补充, 从而对数值信息进行抽取. 随着能够自动抽取特征的深度学习模型的发展, 如双向长短期记忆网络 (bidirectional long short-term conditional random

field, BiLSTM) 模型被研究者所关注, 王竣平等[15] 通过建立数值信息知识库与模板, 抽取属性值与单位, 并利用 BiLSTM-CRF 模型对工业领域中的数值信息进行抽取. Liu 等[16] 设计了包含相对位置特征、绝对位置特征与词典特征等多种外部特征, 并将其向量化后进行连接送入 BiLSTM-CRF 模型进行建模, 从而识别电子医疗病历中的实体与数量信息, 而后将实体数、数量数、相对位置与绝对距离作为外部特征输入随机森林 (random forest) 模型, 对实体与数量信息进行关联. 但以上研究都未对输入信息进行取舍与重要性区分.

此外, 其他研究者针对可度量数量信息的部分信息如实体进行抽取, 商金秋等[17] 利用正向最大匹配算法与决策树模型从电子病历当中抽取患者发热相关症状及其具体表现并将其进行可视化, 以辅助医生治疗. Hundman 等[18] 开发了一个名为 Marve 的系统, 首先利用 CRF 识别数值与单位, 然后基于规则识别实体. Berrahou 等[19] 则是利用 J48 决策树、支持向量机 (support vector machines)、朴素贝叶斯 (naive Bayes)、判别性多义朴素贝叶斯 (discriminative multinominal naive Bayes) 等多个分类器对科学文档中的单位进行抽取. Zhang 等[20] 通过将字符信息与分词信息融入 BiLSTM-CRF 模型, 提升了临床实体识别的性能. Xu 等[21] 将文档级注意力与 BiLSTM 模型结合, 从 2010 i2b2/VA 数据集当中识别临床命名实体, 相比无注意力机制的 BiLSTM 模型提高了 1.01% 的 $F1$ 值, 证明了注意力机制的有效性. 此外, 为了进一步抽取实体, Zhang 等[22] 在通用领域上提出了 Lattice-LSTM, 通过在字符级抽取特征避免分词错误, 并引入当前字符在外部词典中的匹配词来同时考虑字符信息与词信息. 另外, Zhang 等[11] 将句法依存树中的最短依赖路径 (short dependency paths, SDP) 与 RNN 相结合, 排除无关信息. Lin 等[5] 则是将图神经网络 (graph neural network, GNN) 拓展到知识图谱, 以此预测药物之间的反应 (drug-drug interaction, DDI). Song 等[23] 则是将句法依存树拓展为森林, 实现医学关系抽取. 上述部分研究虽利用了注意力机制与剪枝方法进行重要性的区分, 却并未抽取完整的可度量数量信息.

2 可度量数量信息识别与关联模型

2.1 可度量数量信息识别模型

可度量数量信息识别是将输入语句中的每个字符分别标记为实体、数值、单位与其他, 符合序列标记任务的定义. 因此本文将可度量数量信息识别任务转换为一个标准的序列标记任务. 首先将输入语句编码为$X=\{x_1,x_2,x_3,\cdots,x_m\}$, 其中$x_m\in\mathbb{R}^{d_e}$表示语句$X$的第$m$个字符, d_e表示输入向量的维度. 语句的输出标签为$Y=\{y_1,y_2,y_3,\cdots,y_m\}$, 其中$y_m$表示第$m$个字符所对应的标签. 识别任务的目标是寻找一个函数$f_\theta: X\mapsto Y$, 将输入语句的所有字符映射为对应的标签. 对此本文提出 RPA-GRU 模型, 具体模型结构如图 2. 模型首先为输入序列生成对应的向量表示并利用 BiGRU 模型抽取上下文特征, 然后将相对位置向量融入注意力机制对上下文特征进行更新, 以此区分实体与数量信息与非实体和非数量信息, 最后送入 CRF.

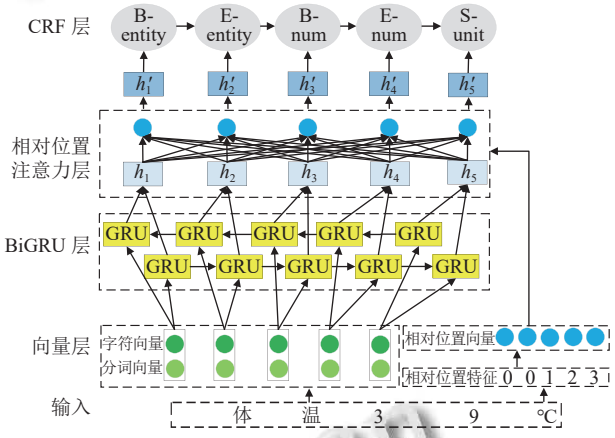


图 2 RPA-GRU 模型的网络结构

(1) 相对位置特征及向量

为了将实体与数值信息与非实体和非数值信息进行区分, 本文对 Liu 等[16] 提出的相对位置特征进行拓展. 具体而言, 对于实体与数量信息, 以距离最近的实体为中心按照距离分配不同的相对位置特征, 对于非实体与非数量信息的相对位置特征而言, 为了防止与实体和数量信息的相对位置特征之间的干扰, 统一设置为语句最大长度+1 之和的负数, 抽取过程如算法 1, 示例为表 1.

表 1 相对位置特征示例

字符	体	温	3	9	℃
相对位置特征	0	0	1	2	3

算法 1. 相对位置特征抽取

输入: 电子病历语句$sen=\{x_1,x_2,\cdots,x_m\}$, sen 中包含的实体与数量信息y, y中的实体 *entity*, sen 的最大长度 *max_len*

输出: 相对位置特征$rf=\{rf_1,rf_2,\cdots,rf_m\}$

```
1) For i=1,⋯, m do
2)   If x_i in y
3)     If x_i in entity
4)       rf_i←0
5)     Else
6)       distance=x_i与最近的 entity 之间的距离
7)     If x_i在距离最近的 entity 左边
8)       rf_i← −1×distance
9)     Else
10)      rf_i← distance
11)    End If
12)  End If
13)  Else
14)    rf_i←−1×(max_len+1)
15)  End If
16) End For
```

本文对相对位置特征进行随机初始化, 并在训练期间进行更新. 从而为输入语句$X=\{x_1, x_2, \cdots, x_m\}$生成对应的相对位置向量$e^{\text{rp}}=\{e_1^{\text{rp}}, e_2^{\text{rp}}, \cdots, e_m^{\text{rp}}\}$.

(2) 相对位置特征融入注意力机制

本文通过将输入语句中的每个字符对应的字符向量与分词向量进行拼接得到$e=[e^{\text{ch}}:e^{\text{seg}}]$作为 BiGRU 模型的输入, 其中$e^{\text{ch}}$与$e^{\text{seg}}$分别为字符向量与分词向量, [:]表示拼接操作. 字符向量由 Word2Vec[24] 进行初始化, 分词向量与相对位置向量类似, 随机初始化后于训练期间更新. 将e送入 BiGRU 模型得到上下文特征$H=[h_1, h_2, \cdots, h_m]$, 从而引入字符与分词信息, 然后将相对位置向量融入注意力机制[25] 中, 为不同部分分配不同重要性, 进一步捕获信息. 计算方式如式 (1):

$$h_x'=\alpha_x h_x,\ x\in\{1,\cdots,m\} \tag{1}$$

其中, α_x为注意力权重, 计算方式如式 (2):

$$\alpha_x=\frac{\exp\left(s(h_x, e_x^{\text{rp}})\right)}{\sum_{n\in\{1,\cdots,m\}}\exp\left(s(h_x, e_n^{\text{rp}})\right)} \tag{2}$$

其中, s为得分函数, 计算方式如式 (3):

$$s(h_x, e_x^{\text{rp}})=v^{\text{T}}\tanh(W_1 h_x+W_2 e_x^{\text{rp}}) \tag{3}$$

通过融入相对位置向量的注意力机制, 得到更新后的上下文特征$H'=[h_1', h_2', \cdots, h_m']$, 最后将$H'$送入标准 CRF 得到最终结果.

2.2 可度量数量信息关联模型

对于需要抽取可度量数量信息的语句而言, 如果单条语句中只有一个可度量数量信息, 那么直接将实体与数量进行关联即可, 然而如图 1, 单条语句中可能存在多个可度量数量信息, 因此需要将语句中的实体与相应的数量进行正确关联. 又由于实体与数量之间仅存在有关联与无关联两种关系, 因此本文将关联任务视作二分类问题. 对此本文提出 GATM 模型, 对实体与数量进行关联, 具体模型结构如图 3. 模型首先将输入语句转换为词向量并生成对应的句法依存树, 对句法依存树进行重构后转换为邻接矩阵, 然后将词向量送入 BiLSTM 获取上下文特征, 将上下文特征与邻接矩阵送入图注意力网络进一步抽取特征, 最后送入 Softmax 得到最终结果.

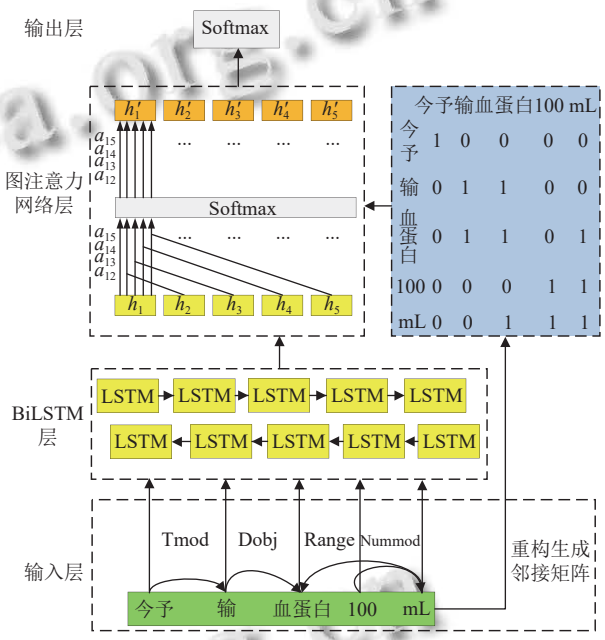
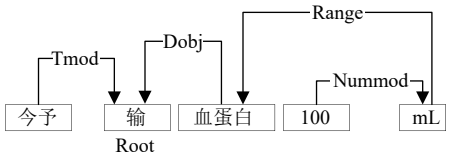


图 3 GATM 模型的网络结构

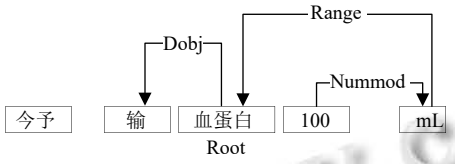
(1) 句法依存树生成与重构

给定输入语句$X=\{x_1, x_2, \cdots, x_l\}$, 其中$l$表示当前输入语句长度. 以输入语句“今予输血蛋白 100 mL”为例, 生成的完整句法依存树示例如图 4(a). 可以看到当前句法依存树根节点为“输”, “tmod”表示时间修饰语, “dobj”表示直接宾语, “range”表示数量词间接宾语, “nummod”表示数词修饰语. 句法依存树描述了各个词语之间的语法联系, 包含着丰富的语义信息, 另外对句法依存树进行适当修剪有助于模型性能的提升. Xu 等[26] 提出基于 SDP 的 LSTM 模型, 通过去除无关信息仅保留两个实体之间的关键路径提升模型的 $F1$ 值. Wang 等[10] 在基于单向 SDP 的基础上提出了双向 SDP (bi-directional SDP) 进一步抽取信息. 另外, 由于本文关心的重点是可度量数量信息但句法依存树通常不以可度量数量信息为根. 因此本文对句法依存树进行以可度

量数量信息中的实体为根的重构, 在重构的同时对句法依存树进行剪枝, 防止无关信息干扰. 重构后的句法依存树如图 4(b), 重构后的句法依存树被转换为邻接矩阵A, $A_{ij}=A_{ji}=1$表示词i与词j在句法依存树中存在依赖关系. 重构过程如算法 2.



(a) 生成的完整句法依存树



(b) 重构后的句法依存树

图 4 句法依存树示例

算法 2. 重构句法依存树

输入: 包含可度量数量信息的语句$sen=\{w_1,w_2,\cdots,w_l\}$, 可度量数量信息中的实体 *ent*, 数量 *quantity*, 原始句法依存树T与直接依赖关系r
输出: 重构后以实体为中心的句法依存树$\widehat{T}$

1) 将 *ent* 作为$\widehat{T}$的根节点
2) **For** $i=1,\cdots,l$ **do**
3) **If** w_i与 *ent* 或 *quantity* 在T中存在直接依赖关系r
4) 向$\widehat{T}$中添加w_i与 *ent* 或 *quantity* 的直接依赖关系r
5) **End If**
6) **End For**

同时为了利用 BiLSTM 模型抽取上下文特征, 本文利用 Word2Vec[24] 将输入语句$X=\{x_1,x_2,\cdots,x_l\}$中的每个词x_i转换为相应的词向量w_i, 从而得到输入语句所对应的词向量序列$W=\{w_1,w_2,\cdots,w_l\}$, 并送入 BiLSTM 模型进行抽取得到相应的上下文特征$H=\{h_1,h_2,\cdots,h_l\}$.

(2) 图注意力网络

GAT 由 Velickovic 等[27] 提出, 其结合了注意力机制与图卷积网络 (graph convolutional network, GCN), 利用注意力机制为不同节点分配不同重要性. 本文将上下文特征H与邻接矩阵A输入 GAT, 得到更新后上下文特征$H'=\{h'_1,h'_2,\cdots,h'_l\}$. 然后将$H'$通过一个线性层, 作为 *Softmax* 层输入, 得到预测向量y, 计算公式如式 (4):

$$y=Softmax(WH'+b)) \tag{4}$$

其中, y为当前输入属于每个类别的概率, 并利用 argmax 函数将其中最大概率的类别作为最终输出. 交叉熵函数作为 GATM 模型的损失函数, 计算方式如式 (5):

$$loss=-\frac{1}{2}\sum_{i=1}^{2}\log(y_i) \tag{5}$$

3 实验

3.1 数据集

实验数据来自某三甲医院烧伤科的 1 359 份电子病历, 最初由两名相关研究人员利用标注工具 Colabeler 对每个句子中的可度量数量信息进行标注, 对于两名研究人员标注不一致的数据, 由一名医学信息学的博士进行最终的标注判定, 并通过 Kappa 检验, 得到最终的实验数据集. 识别数据集格式为 BIOES 标注模式, 其中 B 为 Begin 的缩写, 表示该字符处于开始位置, I 为 Inside 的缩写, 表示该字符处于中间位置, E 为 End 的缩写, 表示该字符处于结束位置, S 为 Single 的缩写, 表示该字符单独构成实体、数值或单位, O 为 Other 的缩写, 表示非实体、非数值与非单位. 数据集具体示例如表 2, 其中 Entity、Num 和 Unit 分别表示实体、数值与单位, “<e></e>”标识当前实体, “<q></q>”标识当前数量. “Entity-Quantity(e, q)”为正例, 表示当前实体与当前数量之间有关联, “Other”为负例, 表示当前实体与当前数量之间无关联.

表 2 数据集具体示例

识别模型	字符	体	温	3	9	℃
	标签	B-Entity	E-Entity	B-Num	E-Num	S-Unit
关联模型	语句	<e>收缩压</e> <q>120 mmhg </q> 舒张压70 mmhg		<e>收缩压</e> 120 mmhg舒张压 <q>70 mmhg</q>		
	标签	Entity-Quantity(e, q)		Other		

最终标注好的数据被随机划分为训练集、验证集与测试集, 数据集详细统计信息如表 3 所示.

表 3 数据集详细统计信息

模型	数据类别	训练集	验证集	测试集	总和
识别模型	语句数	4 730	286	2 809	7 825
	实体数	3 854	210	2 157	6 221
	数值数	3 856	205	2 132	6 193
	单位数	3 701	194	2 032	5 927
关联模型	正例数	1 539	357	207	2 103
	负例数	798	311	127	1 236
	总数	2 337	668	334	3 339

3.2 评价指标

对于识别与关联任务, 本文采用精确率 (*Precision*), 召回率 (*Recall*) 与 *F*1 值作为评价指标, 具体计算方式如式 (6)–式 (8).

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{8}$$

其中, TP表示将正类预测为正类的数量, FP表示将负类预测为正类的数量, FN表示将正类预测为负类的数量.

3.3 基线模型

为了验证 RPA-GRU 在识别任务上的有效性, 本文使用以下基线进行性能比较.

Extended BiLSTM-CRF: Liu 等[16] 将绝对位置特征、相对位置特征与词典特征向量化后进行连接送入 Bi-LSTM-CRF 模型, 提升模型 *F*1 值.

Lattice-LSTM: Zhang 等[22] 利用外部词典匹配句子中的字符, 从而获得包含字符的词语, 生成包含字符与词的格, 从而增强基于字符的模型.

WC-LSTM: Liu 等[28] 分别利用最长单词优先 (longest word first, LWF)、最短单词优先 (shortest word first, SWF)、均值 (average) 与自注意力 (self-attention, SA) 4 种方法在输入的字符向量中融入词汇信息.

LR-CNN: Gui 等[29] 在卷积神经网络 (convolution neural network, CNN) 的基础上利用 Rethinking 机制合并词汇信息, 对匹配语句的字符与潜在单词进行建模.

Soft-Lexicon: Ma 等[30] 通过将每个字符所对应的全部词进行合并后进行加权求和, 得到词向量并与字符向量进行拼接, 引入词汇信息.

另外, 为了验证 GATM 模型在关联任务上的效果, 与以下基线进行比较.

AGGCN: Guo 等[31] 将完整的句法依存树送入 GCN 当中, 并通过注意力机制实现软剪枝, 此外其在 AGGCN 模型的基础上, 利用 LSTM 模型捕获上下文特征从而提出 C-AGGCN 模型.

Att-BiLSTM: Zhou 等[32] 将注意力机制引入 BiLSTM 模型当中, 探究注意力机制对模型的提升.

PA-LSTM: Zhang 等[33] 在 LSTM 模型的基础上引入位置注意力来考虑实体的全局位置信息.

3.4 实验参数

为防止 RPA-GRU 与 GATM 模型产生过拟合, 本文在训练过程中引入正则化, 另外将 Adam[34] 与 AdaGrad[35] 分别作为 RPA-GRU 与 GATM 模型的优化器, 其余参数设置如表 4.

表 4 模型的实验参数设置

参数	RPA-GRU	GATM
学习率	0.005	0.1
字符向量维度	50	N/A
分词向量维度	50	N/A
相对位置向量维度	50	N/A
词向量维度	N/A	100
隐藏层维度	100	100
正则化	0.5	0.3
优化器	Adam	AdaGrad
批次大小	128	50
迭代次数	100	100

3.5 实验结果

模型在识别任务上的实验结果如表 5, 结果表明 RPA-GRU 模型取得了 98.56% 的精确率, 96.61% 的召回率, 97.58% 的 *F*1 值, 在 3 个指标上均超越了其他基线模型. 具体而言, 与之前将外部特征向量化后并连接送入 BiLSTM 模型的 Extended Bi-LSTM-CRF 模型相比, RPA-GRU 模型取得的 *F*1 值高 3.31%, 证明比起简单的特征拼接, 本文将相对位置注意力融入注意力机制更新上下文特征取得的效果更优. 与之前通过外部词典来引入词信息的模型 Lattice-LSTM、WC-LSTM (LWF/SWF/Average)、WC-LSTM (SA)、LR-CNN、Soft-Lexicon 对比, RPA-GRU 模型取得的 *F*1 值分别高 2.30%、2.30%、2.17%、3.10%、2.62%, 证明本文所提出的模型即使不依赖外部词典获取词信息也能获得更好的性能.

表 5 识别任务实验结果对比 (%)

模型	*Precision*	*Recall*	*F*1
Extend BiLSTM-CRF	93.81	94.74	94.27
Lattice-LSTM	96.05	94.53	95.28
WC-LSTM (LWF/SWF/Average)	94.63	95.93	95.28
WC-LSTM (SA)	94.45	96.39	95.41
LR-CNN	92.97	96.04	94.48
Soft-Lexicon	94.67	95.25	94.96
RPA-GRU	98.56	96.61	97.58

模型在识别任务上的混淆矩阵如图 5 所示. 由于混淆矩阵中的实际标签 Other 被预测为 Other 的数量对模型性能没有影响, 因此为简化矩阵, 将其数量置为

0. 从混淆矩阵中可以看到, 对于 Entity、Num 和 Unit 而言, 大部分相关信息都已被成功抽取, 且彼此之间很少发生混淆, 得到了不错的效果, 然而无论是 Entity、Num 还是 Unit 都会与其他信息之间发生一定的混淆, 如“D-二聚体”等实体还是难以进行准确抽取, 导致模型性能受到些许影响.

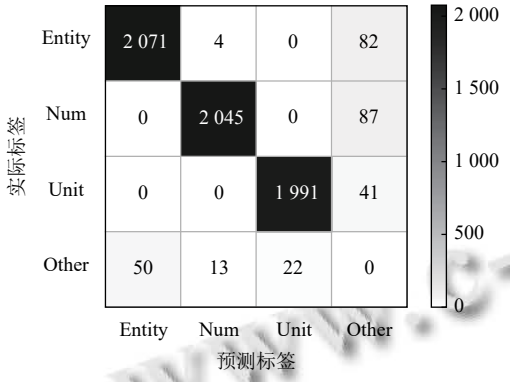


图 5 识别任务混淆矩阵

模型在关联任务上的实验结果如表 6 所示, 结果表明 GATM 模型取得了 96.26% 的精确率, 99.52% 的召回率与 97.86% 的 *F*1 值, 在 3 个指标上均超越了其他基线模型. 具体而言, 与之前利用注意力机制的软剪枝方法 (如 AGGCN 与 C-AGGCN) 相比, GATM 模型高 3.52% 与 2.60% 的 *F*1 值, 证明本文针对句法依存树的重构策略更优. 与仅引入注意力机制的模型如 (Att-BiLSTM、PA-LSTM) 相比, GATM 模型高 3.42% 与 1.74% 的 *F*1 值, 表明 GATM 模型通过引入句法依存树中的句法信息, 有效提升了模型性能.

表 6 关联任务实验对比 (%)

模型	*Precision*	*Recall*	*F*1
C-AGGCN	93.49	97.10	95.26
AGGCN	92.17	96.62	94.34
Att-BiLSTM	95.90	93.03	94.44
PA-LSTM	94.66	97.64	96.12
GATM	96.26	99.52	97.86

模型在关联任务上的混淆矩阵如图 6 所示. 从混淆矩阵中可以看到, 得到的最终结果较为理想, 未发生大规模的混淆情况, 进一步验证了模型的有效性.

随着迭代次数的不断增加, RPA-GRU 模型与 GATM 模型的准确率与损失函数曲线分别如图 7 与图 8 所示. 可以看到, 两个模型的准确率逐步上升, 而损失函数的值逐步减少, 最终都趋于稳定.

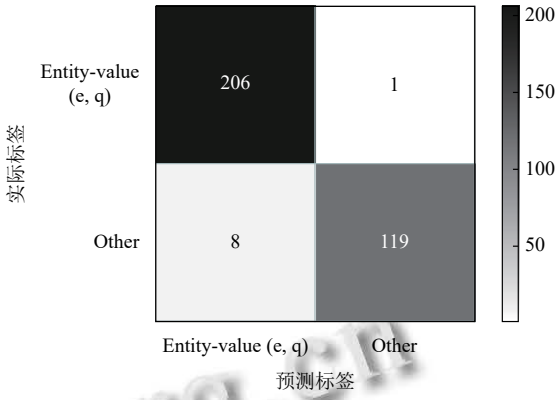


图 6 关联任务混淆矩阵

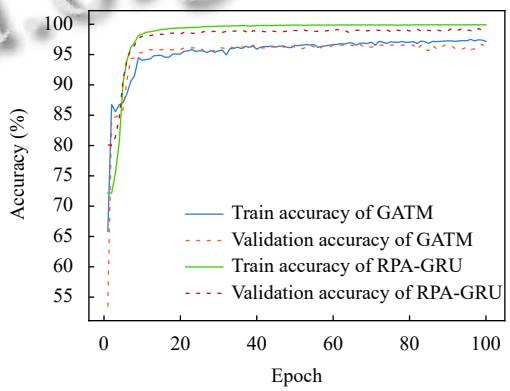


图 7 准确率变化曲线

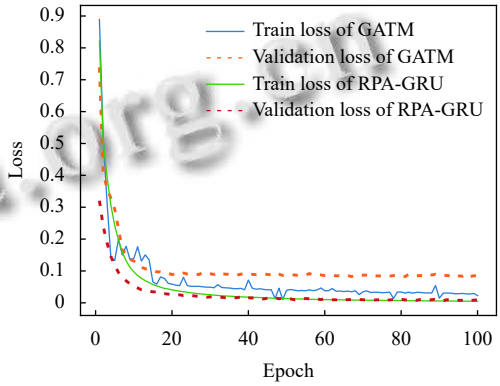


图 8 损失函数变化曲线

为了分析不同训练集大小对 RPA-GRU 模型与 GATM 模型性能的影响, 本文通过随机抽取的方法设置 6 个不同规模大小的训练集, 数据集大小分别原始数据集的 0.10、0.15、0.25、0.50、0.75、1.00. 图 9 显示了在不同训练集大小上训练得到模型的 *F*1 值, 从图中可以看到当训练集大小占比小于 0.25 时, 随着训练集大小的增加, RPA-GRU 模型与 GATM 模型的性能均有着显著的提升, 当训练集大小超过 0.25 时, RPA-

GRU 模型逐渐稳定, GATM 模型则是在训练集大小达到 0.75 时逐渐稳定.

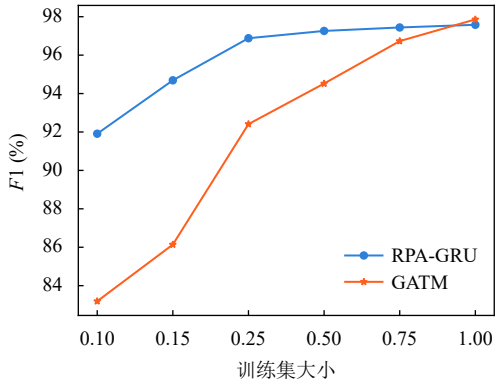


图 9 不同训练集大小的模型性能

4 结论与展望

本文通过对可度量数量信息进行识别与关联完成对于可度量数量信息的抽取, 分别提出了 RPA-GRU 模型与 GATM 模型, 其中 RPA-GRU 模型将相对位置特征融入注意力机制, 对上下文特征进行更新, 有效地提高了模型的性能, 达到了 97.58% 的 *F*1 值. GATM 模型则是以可度量数量信息中的实体为中心重构句法依存树并排除无关信息干扰, 最终取得了 97.86% 的 *F*1 值. 与其他基线模型对比两个模型均取得了最优性能, 证明了其有效性. 此外, 本文还对模型的稳定性进行了探究, 结果证明 RPA-GRU 模型与 GATM 模型在对应的任务中具有稳定的性能.

参考文献

1 Yang X, Bian J, Hogan WR, *et al*. Clinical concept extraction using transformers. Journal of the American Medical Informatics Association, 2020, 27(12): 1935–1942. [doi: 10.1093/jamia/ocaa189]

2 Shi X, Yi YP, Xiong Y, *et al*. Extracting entities with attributes in clinical text via joint deep learning. Journal of the American Medical Informatics Association, 2019, 26(12): 1584–1591. [doi: 10.1093/jamia/ocz158]

3 Hao TY, Pan XY, Gu ZY, *et al*. A pattern learning-based method for temporal expression extraction and normalization from multi-lingual heterogeneous clinical texts. BMC Medical Informatics and Decision Making, 2018, 18(S1): 22. [doi: 10.1186/s12911-018-0595-9]

4 Wei Q, Ji ZC, Li ZH, *et al*. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. Journal of the American Medical Informatics Association, 2020, 27(1): 13–21. [doi: 10.1093/jamia/ocz063]

5 Lin X, Quan Z, Wang ZJ, *et al*. KGNN: Knowledge graph neural network for drug-drug interaction prediction. Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama: ACM, 2021. 380. [doi: 10.5555/3491440.3491820]

6 Hao TY, Wei YY, Qiang JQ, *et al*. The representation and extraction of qunatitative information. Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13). Montperllier: ACL, 2017. 74–83.

7 Hao TY, Liu HF, Weng CH. Valx: A system for extracting and structuring numeric lab test comparison statements from text. Methods of information in Medicine, 2016, 55(3): 266–275. [doi: 10.3414/ME15-01-0112]

8 Liu SS, Pan XY, Chen BY, *et al*. An automated approach for clinical quantitative information extraction from Chinese electronic medical records. Proceedings of the 7th International Conference on Health Information Science. Cairns: Springer, 2018. 98–109. [doi: 10.1007/978-3-030-01078-2_9]

9 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

10 Wang HL, Qin K, Lu GM, *et al*. Direction-sensitive relation extraction using Bi-SDP attention model. Knowledge-Based Systems, 2020, 198: 105928. [doi: 10.1016/j.knosys.2020.105928]

11 Zhang YJ, Zheng W, Lin HF, *et al*. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics, 2018, 34(5): 828–835. [doi: 10.1093/bioinformatics/btx659]

12 肖洪, 薛德军. 基于大规模真实文本的数值知识元挖掘研究. 计算机工程与应用, 2008, 44(30): 150–152, 222. [doi: 10.3778/j.issn.1002-8331.2008.30.046]

13 Turchin A, Kolatkar NS, Grant RW, *et al*. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. Journal of the American Medical Informatics Association, 2006, 13(6): 691–695. [doi: 10.1197/jamia.M2078]

14 张桂平, 张宁, 白宇. 面向问答的数值信息抽取. 郑州大学学报 (理学版), 2018, 50(4): 21–25, 30. [doi: 10.13705/j.issn.1671-6841.2017307]

15 王竣平, 白宇, 蔡东风. 采用 BI-LSTM-CRF 模型的数值信息抽取. 计算机应用与软件, 2019, 36(5): 138–144. [doi:

10.3969/j.issn.1000-386x.2019.05.025]
16 Liu SS, Nie WJ, Gao DF, *et al*. Clinical quantitative information recognition and entity-quantity association from Chinese electronic medical records. International Journal of Machine Learning and Cybernetics, 2021, 12(1): 117–130. [doi: 10.1007/s13042-020-01160-0]
17 商金秋, 朱卫国, 樊银亭, 等. 基于电子病历可视分析的临床诊断模型. 计算机系统应用, 2016, 25(12): 100–107. [doi: 10.15888/j.cnki.csa.005465]
18 Hundman K, Mattmann CA. Measurement context extraction from text: Discovering opportunities and gaps in earth science. arXiv: 1710.04312, 2017.
19 Berrahou SL, Buche P, Dibie-Barthelemy J, *et al*. How to extract unit of measure in scientific documents? Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing. Vilamoura: KDIR, 2013. 249–256. [doi: 10.5220/0004666302490256]
20 Zhang Y, Wang XW, Hou Z, *et al*. Clinical named entity recognition from Chinese electronic health records via machine learning methods. JMIR Medical Informatics, 2018, 6(4): e50. [doi: 10.2196/medinform.9965]
21 Xu GH, Wang CY, He XF. Improving clinical named entity recognition with global neural attention. Proceedings of the 2nd International Joint Conference on Web and Big Data. Macao: Springer, 2018. 264–279. [doi: 10.1007/978-3-319-96893-3_20]
22 Zhang Y, Yang J. Chinese NER using lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 1554–1564. [doi: 10.18653/v1/P18-1144]
23 Song LF, Zhang Y, Gildea D, *et al*. Leveraging dependency forest for neural medical relation extraction. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 208–218. [doi: 10.18653/v1/D19-1020]
24 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. 1st International Conference on Learning Representations. Scottsdale: ICLR, 2013. 1–12.
25 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations. Sab Diego: ICLR, 2015. 1–15.
26 Xu Y, Mou LL, 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: ACL, 2015. 1785–1794. [doi: 10.18653/v1/D15-1206]
27 Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. 6th International Conference on Learning Representations. Vancouver: ICLR, 2018. 1–12.
28 Liu W, Xu TG, Xu QH, *et al*. An encoding strategy based word-character LSTM for Chinese NER. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 2379–2389. [doi: 10.18653/v1/N19-1247]
29 Gui T, Ma RT, Zhang Q, *et al*. CNN-based Chinese NER with Lexicon Rethinking. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: IJCAI, 2019. 4982–4988. [doi: 10.24963/ijcai.2019/692]
30 Ma RT, Peng ML, Zhang Q *et al*. Simplify the usage of lexicon in Chinese NER. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL, 2020. 5951–5960. [doi: 10.18653/v1/2020.acl-main.528.]
31 Guo ZJ, Zhang Y, Liu W. Attention guided graph convolutional networks for relation extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 241–251. [doi: 10.18653/v1/P19-1024]
32 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: ACL, 2016. 207–212. [doi: 10.18653/v1/P16-2034]
33 Zhang YH, Zhong V, Chen DQ, *et al*. Position-aware attention and supervised data improve slot filling. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 35–45. [doi: 10.18653/v1/D17-1004]
34 Kingma DP, Ba LJ. Adam: A method for stochastic optimization. International Conference on Learning Representations. San Diego: ICLR, 2015. 13.
35 Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 2011, 12(61): 2121–2159.

(校对责编: 孙君艳)