第 46 卷　第 6 期　　计　算　机　工　程　　2020 年 6 月
Vol. 46　No. 6　　Computer Engineering　　June 2020

·开发研究与工程应用·　　文章编号：1000-3428(2020)06-0296-07　　文献标志码：A　　中图分类号：TP391

融合双向 GRU 与注意力机制的医疗实体关系识别

张志昌，周　侗，张瑞芳，张敏钰

（西北师范大学 计算机科学与工程学院，兰州 730070）

摘　要： 传统的实体关系识别方法多数是以单个句子作为处理单元，难以解决训练语料中实体关系标签标注错误的问题，且没有充分利用包含实体信息的多个句子在分类实体关系时的相互增强作用。为此，提出一种双向门控循环单元（GRU）和双重注意力机制结合的中文电子病历医疗实体关系识别方法。构建 BiGRU-Dual Attention 模型，采用双向 GRU 学习字的上下文信息，以获取更细粒度的特征，通过引入字级注意力机制提高对关系识别起决定作用的字权重，同时利用句子级注意力机制从多个句子中获取可增强识别性能的特征，降低标注错误的句子对分类的影响。实验结果表明，与 BiLSTM-Attention 模型相比，该模型的 *F*1 值提高了 3.97%，达到了 82.17%。

关键词： 中文电子病历；医疗实体关系抽取；双向门控循环单元；双重注意力机制；深度学习

开放科学（资源服务）标志码（OSID）：

中文引用格式： 张志昌，周侗，张瑞芳，等. 融合双向 GRU 与注意力机制的医疗实体关系识别[J]. 计算机工程，2020，46(6)：296-302.

英文引用格式： ZHANG Zhichang，ZHOU Tong，ZHANG Ruifang，et al. Medical entity relation recognition combining bidirectional GRU and attention[J]. Computer Engineering，2020，46(6)：296-302.

Medical Entity Relation Recognition Combining Bidirectional GRU and Attention

ZHANG Zhichang，ZHOU Tong，ZHANG Ruifang，ZHANG Minyu

（School of Computer Science and Engineering，Northwest Normal University，Lanzhou 730070，China）

【Abstract】 Most of existing methods for entity relationship recognition take a single sentence as processing unit，and fail to address tagging errors of entity relationships in the training corpus. Also，they cannot make full use of the mutual reinforcement of multiple sentences that contain entity information in relationship recognition. Therefore，this paper proposes a recognition method based on bidirectional Gated Recurrent Unit（GRU）and dual attention mechanism for entity relationships of Chinese electronic medical records. This paper proposes a BiGRU-Dual Attention model，and uses bidirectional GRU to learn the context information of characters in order to obtain more fine-grained features. Then the character-level attention mechanism is introduced to improve the weight of the characters that are key to relation recognition. Also，the sentence-level attention mechanism is employed to capture the features that can enhance recognition performance from multiple sentences，so as to reduce the weight of mislabeled sentences. Experimental results show that compared with the mainstream BiLSTM-Attention model，the proposed model increases the *F*1 value by 3.97% to 82.17%.

【Key words】 Chinese Electronic Medical Records（EMR）；medical entity relation extraction；bidirectional Gated Recurrent Unit（GRU）；dual attention mechanism；deep learning

DOI： 10.19678/j.issn.1000-3428.0054431

0　概述

电子病历（Electronic Medical Records，EMR）是医务人员使用电子医疗系统产生的文字、符号、图表、图形、数据和影像等数字化信息，并将其进行存储的医疗记录[1]。随着 EMR 的大量使用，人们对其认识也逐渐完善，它不仅包括患者的一些临床信息，如检查结果、临床诊断以及不良反应等，还包括丰富的医疗实体[2]。如何在非结构化的病历文本中抽取有价值的医疗信息，建立可用于临床决策支持的医

基金项目： 国家自然科学基金（61762081，61662067，61662068）；甘肃省重点研发计划项目（2017GS10781）。

作者简介： 张志昌（1976—），男，教授、博士，主研方向为医疗文本处理、问答技术；周　侗、张瑞芳、张敏钰，硕士研究生。

收稿日期： 2019-03-28　　**修回日期：** 2019-06-03　　**E-mail：** zzc@nwnu.edu.cn

疗知识库,成为自然语言处理(Natural Language Processing,NLP)领域的研究热点。实体关系抽取是NLP信息抽取技术中的基本任务,也是构建知识库和知识图谱的关键方法[3]。从EMR文本中挖掘医疗实体以及实体间的语义关系,对于推动EMR在医疗健康服务中的应用具有重要意义。实体关系抽取最早被消息理解会议(Message Understanding Conference,MUC)[4]评测会议引入,直至2010年,I2B2/VA在NLP挑战临床记录中提出关于英文EMR的医疗实体关系抽取[5],使得EMR中的医疗实体关系抽取成为了研究热点。但在中文EMR方面,公开的评测以及研究成果相对较少,已有的关系抽取方法依赖于机器学习算法,且需要构建大量的手工特征。近年来,在不依赖手工特征条件下,神经网络方法在关系抽取任务中取得了较好的性能,但是常见的关系抽取是以句子作为单独的处理单元,没有考虑到EMR语料库中部分语料的实体关系标签标注错误,影响分类效果。

本文提出一种双向门控循环单元(Gated Recurrent Unit,GRU)和双重注意力机制结合的深度学习方法。该方法构建一个双向GRU和双重注意力机制结合的实体关系抽取模型,利用双向GRU学习字的上下文信息,获取更细粒度的特征。通过字级注意力机制提高对关系分类起决定作用的字权重,利用句子级注意力机制学习更多语句的特征,降低噪声句子的权重,以有效解决标签标注错误问题,提高分类器效果。

1　相关研究

目前,大多数关于实体关系抽取的方法是在开放域上进行的,如新闻报道、博客以及维基百科等[6]。在开放域上进行关系抽取研究的最大难点在于语料内容没有固定的结构,早期的实体关系抽取研究是基于有监督学习的方法,如基于特征工程、核函数以及条件随机场[7]的方法。文献[8]在MUC-7评测会议中,对原始数据进行统计并提取特征来进行实体关系抽取,实验过程中取得了较高的$F1$值。文献[9]利用支持向量机的方法进行关系抽取,这类方法依赖于人工构建手工特征,需要标注大量的训练语料,耗时耗力,且泛化能力差。针对此局限性,文献[10]提出远程监督的思想,通过将文本与大规模知识图谱进行实体对齐,有效解决关系抽取的标注数据规模问题。文献[11]首先使用循环神经网络来解决关系抽取问题,利用句法结构得到句子的向量表示并用于关系分类,但没有考虑到实体在句子中的位置和语义信息。文献[12]利用卷积神经网络进行关系抽取,采用词向量和词位置向量作为输入,通过卷积、池化得到句子表示,使得在关系抽取过程中考虑到句子中的实体信息。文献[13]提出一种基于最短依存路径表示文本的深度学习方法,能够准确地抽取实体关系。

医疗领域的关系抽取与开放域的关系抽取有所不同,EMR是一种半结构化的文本数据,包含大量的专业术语、缩略词等。2010年,I2B2/VA评测引入了英文EMR的信息抽取任务[5],定义了三大类医疗实体关系:1)医疗问题和治疗的关系;2)医疗问题和检查的关系;3)医疗问题和医疗问题的关系。文献[14]使用支持向量机作为分类器,并引入外部字典和丰富的特征提升关系识别精度。文献[15]通过基于规则的方法从中草药相关文章中抽取关系,并用于构建关系数据库。文献[16]从病历中计算疾病和症状的共现程度来抽取两者的关系。文献[17]采用两阶段方法,将长短期记忆(Long Short Term Memory,LSTM)网络和支持向量机相结合,抽取药物之间的影响关系。

2　方法描述

给定一个句子集合$S=\{x_1,x_2,\cdots,x_n\}$,其中x_i为句子集合S中的第i个句子。实验模型主要分为句子编码和句子级注意力机制两部分。

2.1　句子编码

句子编码模型如图1所示,将任意给定的一个句子$x_i=\{c_1,c_2,\cdots,c_n\}$通过双向GRU编码处理,字级注意力机制计算产生每个字的权值,并把双向GRU的输出向量表示成一个句子向量。

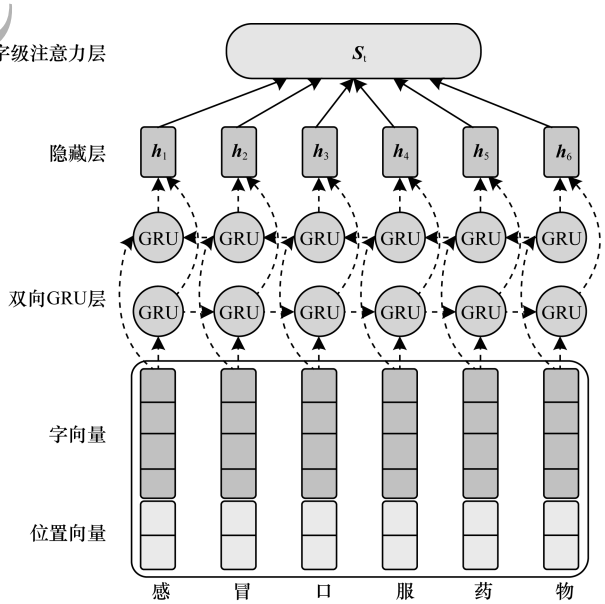

图1　句子编码模型

Fig.1　Model of sentence coding

2.1.1 向量表示

任意给定一个句子 $x_i=\{c_1,c_2,\cdots,c_n\}$ 作为双向 GRU 的输入,其中包含 2 个实体 e_1 和 e_2。将句子中的每个字 c_i 映射到一个低维稠密向量 $\boldsymbol{V}_i=(V_w^i,V_p^i)$,其中,$V_w^i$ 表示字向量,V_p^i 表示字相对实体的位置向量。字向量和位置向量具体描述如下:

1)字向量表示:字嵌入是将句子中的字映射成一个低维稠密的向量,从而更好地刻画字的属性。给定一个含有 n 个字的句子 $x_i=\{c_1,c_2,\cdots,c_n\}$,实验用 word2vec 工具训练生成字向量,每个字均被映射为向量表示,向量维度为 d_w。

2)位置向量表示:在关系抽取任务中,位置嵌入用相对位置的低维向量表示,最早被文献[12]引入实体关系抽取任务中。在图 2 所示标注的句子中,当前字“引”与医疗实体“感冒”“发烧”之间的相对位置分别为 2 和 -2,每个相对位置分别对应一个位置向量,维度为 d_p。

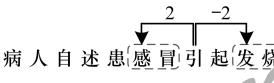

图 2 当前字与医疗实体的相对位置

Fig. 2 Relative position of the current word and the medical enity

最后,实验将字向量和位置向量连接起来,并将其表示为一个向量 $\boldsymbol{V}_i=(\boldsymbol{V}_w^i,\boldsymbol{V}_p^i)$,向量维度为 $d_v=d_w+2d_p$。

2.1.2 双向 GRU 层

GRU 是循环神经网络的分支,也是 LSTM 的变体,GRU 在保持 LSTM 效果的同时使其结构简单,且计算简便,由于其在序列处理上的出色表现而被广泛应用于自然语言处理任务中。GRU 结构如图 3 所示。

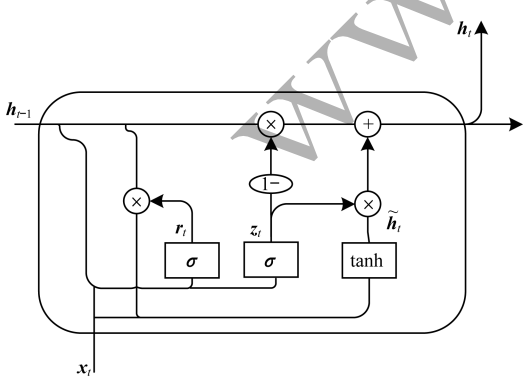

图 3 GRU 结构

Fig. 3 GRU structure

$$\boldsymbol{z}_t=\sigma(\boldsymbol{W}_z\boldsymbol{x}_t+\boldsymbol{U}_z\boldsymbol{h}_{t-1}+\boldsymbol{b}_z) \tag{1}$$

$$\boldsymbol{r}_t=\sigma(\boldsymbol{W}_r\boldsymbol{x}_t+\boldsymbol{U}_r\boldsymbol{h}_{t-1}+\boldsymbol{b}_r) \tag{2}$$

$$\tilde{\boldsymbol{h}}_t=\tanh(\boldsymbol{W}_h\boldsymbol{x}_t+\boldsymbol{U}_h(\boldsymbol{h}_{t-1}\otimes\boldsymbol{r}_t)+\boldsymbol{b}_h) \tag{3}$$

$$\boldsymbol{h}_t=(1-\boldsymbol{z}_t)\otimes\boldsymbol{h}_{t-1}+\boldsymbol{z}_t\otimes\tilde{\boldsymbol{h}}_t \tag{4}$$

$$\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}_t},\overleftarrow{\boldsymbol{h}_t}] \tag{5}$$

其中,z_t 和 $\boldsymbol{r}_t$ 分别为 GRU 的更新门和重置门,更新门是控制上一时刻的状态信息传递到当前时刻的程度,重置门是控制上一时刻的状态信息被遗忘的程度。$\boldsymbol{W}_z,\boldsymbol{W}_r,\boldsymbol{W}_h$ 和 $\boldsymbol{U}_z,\boldsymbol{U}_r,\boldsymbol{U}_h$ 分别为神经元当前时刻的输入权重和循环输入的权重,$\boldsymbol{b}_z,\boldsymbol{b}_r,\boldsymbol{b}_h$ 为偏置向量。首先,实验通过上一时刻的隐藏状态信息 $\boldsymbol{h}_{t-1}$ 和当前时刻的节点输入 $\boldsymbol{x}_t$ 来获取 2 个门控的状态。得到门控信号之后,利用重置门来获取遗忘后的状态 $\boldsymbol{h}_{t-1}\otimes\boldsymbol{r}_t$,⊗表示哈达马积对应元素相乘;然后,将其与当前时刻的输入 $\boldsymbol{x}_t$ 相加并通过非线性函数 tanh 激活;最后,用更新门对当前节点的输入选择记忆。

GRU 采用“门”结构来克服短时记忆的影响,不仅可以调节流经序列的信息流,还可以改善 RNN 存在的“梯度消失”问题。为了能够有效利用上下文信息,实验采用双向 GRU 结构,双向 GRU 对每个句子分别采用前向和反向计算得到 2 个不同的隐藏层状态,然后将 2 个向量相加得到最终的编码表示。

2.1.3 字级注意力机制

注意力机制模仿了生物观察行为的内部过程,是一种通过增加部分区域的注意力来获取关注目标更多细节信息的机制。注意力机制可以快速提取数据的重要特征,减少对外部信息的依赖,捕获语言中的长距离依赖性,被广泛应用于自然语言处理任务中。本文通过引入字级注意力机制来判断每个字对关系分类的重要程度,并有效提高模型精确率。

通过双向 GRU 得到每个字的输出向量 $\boldsymbol{h}_t$,输入到全连接层并获得其隐藏表示 $\boldsymbol{u}_t$,通过 Softmax 函数计算归一化权重向量 $\boldsymbol{\alpha}_t$,最后得到句子向量表示。字级注意力机制权重计算如下:

$$\boldsymbol{u}_t=\tanh(\boldsymbol{h}_t\boldsymbol{W}_t+\boldsymbol{b}_t) \tag{6}$$

$$\boldsymbol{\alpha}_t=\frac{\exp(\boldsymbol{u}_t\boldsymbol{u}_w)}{\sum_{t=1}^{T}\exp(\boldsymbol{u}_t\boldsymbol{u}_w)} \tag{7}$$

$$\boldsymbol{S}_t=\sum_{t=1}^{T}\boldsymbol{\alpha}_t\boldsymbol{h}_t \tag{8}$$

其中,$\boldsymbol{W}_t$ 表示当前时刻神经元的输入权重,T 表示序列长度,$\boldsymbol{u}_w$ 表示随机初始化的上下文向量,通过反向传播更新上下文向量。$\boldsymbol{S}_t$ 表示编码后的句子向量。

2.2　句子级注意力机制

目前,很多用来构建知识库的方法均需要标注好的训练语料,人工标注的语料因为标注人员不同而导致语料噪声。在实验标注的语料库中,相同的实体对和实体类型在不同的语料中被标注为不同的关系标签,影响模型效果。常见的关系抽取方法是以句子作为单独的处理单元,若仅使用字级注意力机制时,则只考虑到当前的句子信息,而对于含有同一实体对的其他句子,还需要通过句子级注意力机制学习实体共现句的上下文特征,为每个句子学习注意力权重,来提升分类器效果。正确标注的句子将获得较高的权重,而错误标注的句子会得到较低的权重,隐式摒弃一些噪声语料,如图 4 所示。

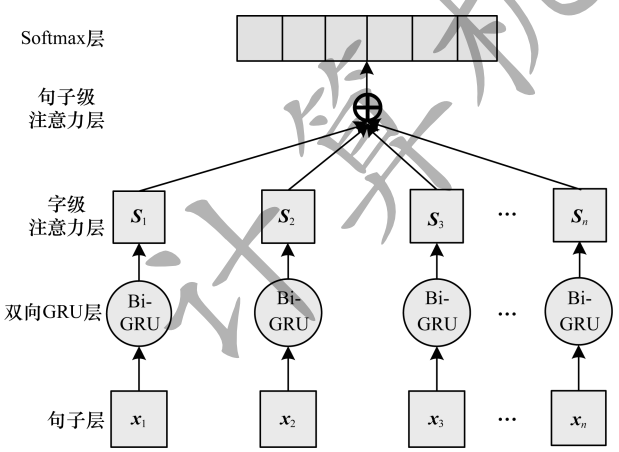

图 4　句子级注意力机制模型

Fig. 4　Model of sentence-level attention mechanism

对于给定的一组实体对(e_1,e_2),所有它们共同出现的 n 个句子组成集合 $\boldsymbol{S}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,句子级注意力机制为该集合计算相对应的权值向量 $\boldsymbol{\alpha}=\{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_n\}$。实验用 $\vec{\boldsymbol{S}}$ 表示集合 $\boldsymbol{S}$ 的向量,它是所有句子向量的加权和,其特征向量计算如下:

$$\vec{\boldsymbol{S}}=\sum_{i=1}^{n}\boldsymbol{\alpha}_i\boldsymbol{x}_i \tag{9}$$

通过计算句子特征向量与目标实体关系的相似度来得到句子的注意力权值。句子特征向量与目标实体关系向量的相似度越高,则正确表达实体关系的可能性越大,注意力权重也越高。句子特征向量目标实体关系的相似度计算如下:

$$\boldsymbol{\alpha}_i=\frac{\exp(e_i)}{\sum_n\exp(e_j)} \tag{10}$$

$$e_i=\boldsymbol{x}_i\boldsymbol{A}\boldsymbol{r} \tag{11}$$

其中,e_i表示句子特征向量 $\boldsymbol{x}_i$与预测关系向量 $\boldsymbol{r}$ 的匹配分数,$\boldsymbol{A}$ 表示加权对角矩阵。最后,通过 Softmax 层对实体关系向量进行输出。

3　实验结果与分析

3.1　数据集

中文 EMR 中包含大量的医学知识和临床信息,由于标注人员医学领域知识的限制以及病历中包含患者的隐私,使得 EMR 在语料构建上存在一定的困难。本文依据 I2B2/VA Challenge 医学关系标注规范,且在专业人员的指导下,制定自己的中文 EMR 标注规范。在 EMR 的文本片段中,医学实体语义关系主要存在于治疗、疾病、检查和症状等实体之间,如表 1 所示,包含 5 个粗粒度类别和 15 个细粒度类别,表 2 所示为标注语料示例。

表 1　医疗实体关系类型及其描述

Table 1　Types and descriptions of medical entity relationships

粗粒度关系	细粒度关系类型	关系类型描述
治疗和疾病的关系	TrID	治疗改善了疾病
	TrWD	治疗恶化了疾病
	TrCD	治疗导致了疾病
	TrAD	治疗施加于疾病
	TrNAD	因为疾病而没有采取治疗
治疗和症状体征的关系	TrIS	治疗改善了症状
	TrWS	治疗恶化了症状
	TrCS	治疗导致了症状
	TrAS	治疗施加于症状
	TrNAS	因为症状而没有采取治疗
检查和疾病的关系	TeRD	检查证实了疾病
	TeCD	为了证实疾病而采取检查
检查和症状体征的关系	TeRS	检查证实了症状
	TeBS	因为症状而采取检查
疾病和症状体征的关系	DCS	疾病导致症状

表 2　中文电子病历医疗实体关系

Table 2　Medical entity relationship of Chinese electronic medical records

中文电子病历示例	实体关系
高血压病口服利血平控制	治疗改善了疾病(TrID)
口服抗生素药物,痰中带血无明显缓解	治疗恶化了症状(TrWS)
头 MRI 示:腔隙性脑梗死	检查证实了疾病(TeRD)
双肺听诊可闻及少量痰鸣音	检查证实了症状(TeRS)
3 年前脑梗死遗留尿便失禁	疾病导致症状(DIS)

本文以甘肃省某二级甲等医院提供的不同临床科室的 EMR 为分析对象。首先,对已校对的 EMR 文本进行简单的去隐私处理,然后,从不同临床科室随机挑选一定量的 EMR 文本进行人工标注。本文总共使用 1 200 份 EMR 文本对实体关系抽取进行研究,其中 800 份 EMR 作为训练集,200 份 EMR 作为开发集,200 份 EMR 作为测试集。

3.2 评价指标

本文利用精确率 P、召回率 R 和 $F1$ 值对中文 EMR 实体关系分类效果进行评价,具体计算公式如下:

$$P=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}}\times 100\% \tag{12}$$

$$R=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}}\times 100\% \tag{13}$$

$$F1=\frac{2PR}{P+R}\times 100\% \tag{14}$$

其中,T_{P} 表示对当前类别识别正确的数目,F_{P} 表示对当前类别识别错误的数目,F_{N} 表示应该识别为当前类别但是没有被识别的数目,$T_{\mathrm{P}}+F_{\mathrm{N}}$ 表示该类别下所有正实例的总数目,$T_{\mathrm{P}}+F_{\mathrm{P}}$ 表示识别出来属于当前类别的总数。分别计算各个类别的精确率 P 和召回率 R,然后以 $F1$ 值作为各个类别整体的评价指标。

3.3 实验设置

选择目前的主流模型 LSTM 作为基线实验,分别和 SVM 模型、CNN 模型、BiLSTM-Attention 模型和 BiGRU-Dual Attention 模型进行对比。

1)SVM 模型:该模型在 SemEval-2010 评测任务中表现最好。文献[18]利用各种手工制定的特征,用 SVM 作为分类器,实验取得了较好的 $F1$ 值。

2)CNN 模型:该模型被文献[19]使用,采用 CNN 编码句子向量,将编码后的结果最大池化,利用 Softmax 函数输出结果。

3)BiLSTM-Attention 模型:该模型由文献[20]提出。利用双向 LSTM 抽取上下文信息,结合注意力机制对词赋予不同的权重,判断每个词对关系分类的重要程度,提高对分类有贡献的词权重,有效提高模型效率。

4)BiGRU-Dual Attention 模型:该模型由本文提出,使用双向 GRU 和双重注意力机制结合来抽取实体关系,通过随机搜索调整在开发集上的超参数,超参数如表 3 所示。

表 3 BiGRU-Dual Attention 模型超参数

Table 3 Hyperparameters of BiGRU-Dual Attention model

超参数	描述	参数值
d_p	字向量维度	100
d_v	位置向量维度	5
β	Batch Size	50
μ	Epoch Num	10
η	Dropout 比率	0.5

模型实验中字向量维度为 100,位置向量的维度为 5,Batch Size 大小为 50,Epoch Num 设置为 10,使用 Adam 优化器进行训练,学习率为 0.000 5,其中 L2 正则化值为 1,Dropout 比率为 0.5。在本文中,将 Dropout 比率与 L2 正则化结合起来以防止过度拟合。

3.4 实验结果

本文提出基于双向 GRU 和双重注意力机制结合的实体关系抽取模型,将擅长学习长期依赖信息的双向 GRU 加入到句子编码阶段中,然后用字级注意力机制提高对关系分类有决定作用的字权重,最后用句子级注意力机制获取更多语句的特征,增大正确标注的句子权重,同时减小错误标注的句子权重。在训练过程中,使用相同的数据、批次大小及迭代次数,分别对 SVM 模型、CNN 模型、LSTM 模型、BiLSTM-Attention 模型和本文模型进行训练,记录训练过程中最高的精确率 P、召回率 R 和 $F1$ 值,具体数据如表 4 所示。

表 4 不同模型进行中文电子病历实体关系识别时的性能比较

Table 4 Performance comparison of different models for recognizing entity relationship of Chinese electronic medical records %

模型	P	R	$F1$
SVM	66.34	63.26	64.76
CNN	63.52	61.57	62.52
LSTM	74.12	75.30	74.71
BiLSTM-Attention	78.86	77.56	78.20
本文模型	81.51	82.84	82.17

3.5 实验分析

根据上述表 4 中的数据,可以看到本文提出的基于双向 GRU 结合双重注意力机制的实体关系抽取方法相比其他方法效果较好,$F1$ 值达到了82.17% 。表 4 中的学习方法可以分为传统机器学习方法和深度学习方法,从实验结果可以看出,深度学习方法普遍优于传统机器学习方法,这是由于传统机器学习方法依赖于大量的手工特征,而 EMR 中文本语料较长,且结构性差,传统机器学习方法无法从病历文本中获得包含的语义和长距离信息。本文提出的 BiGRU-Dual Attention 模型相较于传统的机器学习算法有明显地提高,同时相较于目前主流的 BiLSTM-Attention 模型,$F1$ 值提高了 3.97% 。在表 4 中,可以看出精确率 P 和召回率 R 均得到了大幅提高,这说明本文提出的方法改善了错误标签的问题,同时在对细粒度特征分析中,结果发现 $F1$ 值也提升了很多。双向 GRU 和注意力机制的影响分析如下:

1)双向 GRU 的影响分析。本文模型在句子编码阶段加入双向 GRU 结构,能够很好地学习字的上下文

信息,并提供丰富的特征。由表 4 可以看出,LSTM 的关系抽取模型比普通卷积的效果更好,然而 GRU 作为 LSTM 的变体,它可以像 LSTM 一样,既具备记忆序列特征的能力,又善于学习长距离依赖信息。EMR 文本语料较长,存在许多长依赖语句,卷积神经网络仅靠滑动窗口来获得局部信息,不能学习到长依赖特征。双向 GRU 结构却可以学习到丰富的上下文特征,且效果更佳。

2)注意力机制的影响分析。本文模型中通过加入注意力机制,来判断每个字对关系分类的重要程度,提高分类效果,并且引入句子级别的注意力机制,增大正确标注的句子权重,同时减小错误标注的句子权重。本文实验设计对比了 LSTM 模型、BiLSTM-Attention 模型以及 BiGRU-Dual Attention 模型的实体关系抽取效果。其中,BiLSTM-Attention 模型只使用字级注意力机制,BiGRU-Dual Attention 模型使用了字级注意力机制和句子级注意力机制。从表 4 的实验结果可以看出,加入注意力机制的方法均高于未加注意力机制的方法,其中 BiLSTM-Attention 模型的 $F1$ 值比 LSTM 模型的高 3.49% ,说明加入字级注意力机制有助于实体关系抽取准确率的提升。此外,由表 4 中数据可以看出,BiLSTM-Attention 方法的 $F1$ 值比本文方法要低许多,这可能是因为本文使用的句子级注意力机制学习更多的语句特征,降低错误标注语句的权值,减少噪声句子影响。

3.6　方法验证

实验将本文方法与 Mintz、MultiR、MIML 3 种传统的远程监督方法相比较,具体数据如图 5 所示。

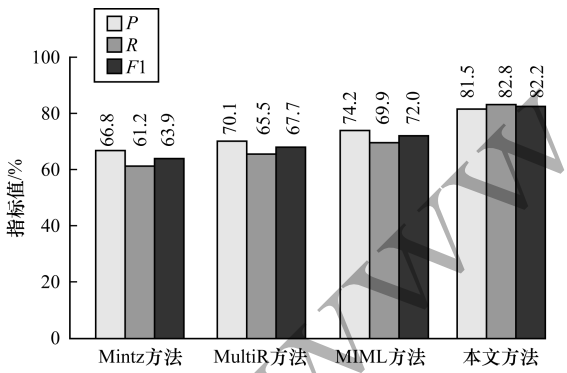

图 5　本文方法与传统远程监督方法的结果对比

Fig. 5　Comparison of the results of proposed method and traditional remote monitoring methods

由图 5 可知,本文方法的精确率 P、召回率 R、$F1$ 值均高于其他 3 种传统的远程监督方法,这是因为本文提出的方法不需要人工构建特征,能够准确学习到句子的语义信息,直接从原始字中自动学习特征,减少错误传播。另外,本文方法除了从更多的语句中学习特征,还加入字级注意力机制和句子级注意力机制,有效缓解标签错误问题。

4　结束语

针对已有实体关系抽取方法存在的标签标注错误问题,本文提出双向 GRU 和双重注意力机制结合的实体关系抽取方法。利用双向 GRU 学习字的上下文信息,获取更细粒度的特征信息,通过字级注意力机制提高对关系分类起决定作用的字权重,同时加入句子级注意力机制学习更多的语句信息,有效解决标签错误问题。通过在人工标注的数据集上进行实验对比,证明了本文方法能有效提升实体关系抽取效果。下一步将对实体识别和实体关系进行联合抽取。

参考文献

[1] Ministry of Health of the People's Republic of China. Basic standards for electronic medical records (trial)[J]. China Health Quality Management, 2010(4): 13-14. (in Chinese)
中华人民共和国卫生部. 电子病历基本规范(试行)[J]. 中国卫生质量管理, 2010(4): 13-14.

[2] SHEN Wei. The change of medical behavior caused by the electronic medical record[J]. Journal of Medical Informatics, 2007, 28(4): 346-347. (in Chinese)
沈伟. 电子病历给医疗行为带来的变革[J]. 医学信息学杂志, 2007, 28(4): 346-347.

[3] ZHANG Xiumei, XU Jianwu, CHENG Yuhua, et al. Knowledge-based building of a clinical decision support system[J]. Chinese Journal of Hospital Administration, 2014, 30(6): 472-475. (in Chinese)
张秀梅, 徐建武, 程煜华, 等. 基于知识库的临床决策支持系统构建[J]. 中华医院管理杂志, 2014, 30(6): 472-475.

[4] GRISHMAN R, SUNDHEIM B. Message understanding conference-6: a brief history[C]//Proceedings of the 16th Conference on Computational Linguistics. New York, USA: ACM Press, 1996: 466-471.

[5] UZUNER Ö, SOUTH B, SHEN S Y, 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.

[6] SARAWAGI S. Information extraction[J]. Foundations and Trends in Databases, 2008, 1(3): 261-377.

[7] ZHOU Jing. Chinese entity relation extraction based on conditional random fields model[J]. Computer Engineering, 2010, 36(24): 192-194. (in Chinese)
周晶. 基于条件随机域模型的中文实体关系抽取[J]. 计算机工程, 2010, 36(24): 192-194.

[8] CHEN H H,DING Y W,TSAI S C,et al. Description of the NTU system used for MET-2[C]//Proceedings of the 7th Message Understanding Conference. San Diego, USA:Internet Society,1998:1-9.

[9] ZHANG Zhu. Weakly-supervised relation classification for information extraction[C]//Proceedings of the 13th ACM Conference on Information and Knowledge Management. New York,USA:ACM Press,2004:581-588.

[10] MINTZ M,BILLS S,SNOW R,et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. New York,USA:ACM Press,2009:1003-1011.

[11] RICHARD S,BRODY H,CHRISTOPHER D,et al. Semantic compositionality through recursive matrix-vector space[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. New York, USA: ACM Press, 2012:1201-1211.

[12] ZENG Daojian, LIU Kang, LAI Siwei, et al. Relation classification via convolutional deep neural network[C]//Proceedings of the 25th International Conference on Computational Linguistics. Dubin, Ireland: [s. n.], 2014: 2335-2344.

[13] SUN Ziyang,GU Junzhong,YANG Jing. Chinese entity relation extraction method based on deep learning[J]. Computer Engineering,2018,44(9):164-170. (in Chinese)
孙紫阳,顾君忠,杨静. 基于深度学习的中文实体关系抽取方法[J].计算机工程,2018,44(9):164-170.

[14] 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.

[15] FANG Y C, HUANG H C, CHEN H H, et al. TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining[J]. BMC Complementary and Alternative Medicine,2008,8(1):58-58.

[16] WANG X Y,CHUSED A,ELHADAD N,et al. Automated knowledge acquisition from clinical narrative reports[C]//Proceedings of the American Medical Informatics Association. New York,USA:ACM Press,2008:783-787.

[17] HUANG Degen,JIANG Zhenchao,ZOU Li,et al. Drug-drug interaction extraction from biomedical literature using support vector machine and long short term memory networks[J]. Information Sciences,2017(415/416):100-109.

[18] RINK B, HARABAGIU S. UTD: classifying semantic relations by combining lexical and semantic resources[C]//Proceedings of the 5th International Workshop on Semantic Evaluation. New York,USA:ACM Press,2010:256-259.

[19] SAHU S K,ANAND A,ORUGANTY K,et al. Relation extraction from clinical texts using domain invariant convolutional neural network[C]//Proceedings of the 15th Workshop on Biomedical Natural Language. Berlin, Germany:Springer,2016:206-215.

[20] LI Lingfeng,NIE Yuanping,HAN Weihong,et al. A multi-attention-based bidirectional long short-term memory network for relation extraction[C]//Proceedings of the International Conference on Neural Information. Berlin,Germany:Springer, 2017:216-227.

编辑 刘继娟