计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.009322]

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

基于知识增强的中文电子病历命名实体识别①

李宛泽, 宋　波, 齐岳山

(青岛科技大学 信息科学技术学院, 青岛 266061)
通信作者: 宋　波, E-mail: songbo@kedauis.com

摘　要: 针对中文电子病历中医疗嵌套实体难以处理的问题, 本文基于 RoBERTa-wwm-ext-large 预训练模型提出一种知识增强的中文电子病历命名实体识别模型 ERBEGP. RoBERTa-wwm-ext-large 采用的全词掩码策略能够获得词级别的语义表示, 更适用于中文文本. 首先结合知识图谱, 使模型学习到了大量的医疗实体名词, 进一步提高模型对电子病历实体识别的准确性. 然后通过 BiLSTM 对电子病历输入序列编码, 能够更好捕获病历的中上下语义信息. 最后利用全局指针网络模型 EGP (efficient GlobalPointer) 同时考虑实体的头部和尾部的特征信息来预测嵌套实体, 更加有效地解决中文电子病历命名实体识别任务中嵌套实体难以处理的问题. 在 CBLUE 中的 4 个数据集上本文方法均取得了更好的识别效果, 证明了 ERBEGP 模型的有效性.

关键词: 中文电子病历; 命名实体识别; 知识增强; 嵌套实体; 全局指针网络模型; 深度学习

引用格式: 李宛泽,宋波,齐岳山.基于知识增强的中文电子病历命名实体识别.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9322.html

Knowledge-enhanced Named Entity Recognition for Chinese Electronic Medical Records

LI Wan-Ze, SONG Bo, QI Yue-Shan

(School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Regarding the challenge of handling nested medical entities in Chinese electronic medical records, this study proposes a knowledge-enhanced named entity recognition model for Chinese electronic medical records called ERBEGP based on the RoBERTa-wwm-ext-large pre-trained model. The comprehensive word masking strategy employed by the RoBERTa-wwm-ext-large model can obtain semantic representations at the word level, which is more suitable for Chinese texts. First, the model learns a significant number of medical entity nouns by integrating knowledge graphs, further improving entity recognition accuracy in electronic medical records. Then, the contextual semantic information within the records can be better captured through BiLSTM encoding of the input sequence of medical records. Finally, the efficient GlobalPointer (EGP) model is adopted to simultaneously consider the features of both the head and tail of entities to predict nested entities, addressing the challenge of handling nested entities in named entity recognition tasks of Chinese electronic medical records. The effectiveness of the ERBEGP model is demonstrated by yielding better recognition results on the four datasets within CBLUE.

Key words: Chinese electronic medical records; named entity recognition (NER); knowledge enhancement; nested entities; global pointer network model; deep learning

1　引言

中文电子病历命名实体识别是医学信息处理和自然语言处理领域的研究热点之一. 近年来, 随着电子病历系统的广泛应用, 医疗领域产生了大量的中文电子

① 收稿时间: 2023-05-22; 修改时间: 2023-06-28; 采用时间: 2023-07-12; csa 在线出版时间: 2023-09-22

病历数据. 这些数据包含了丰富的医学信息, 如疾病名称、药物名称、手术名称等, 因此可以作为医学知识的宝库. 然而, 中文电子病历数据规模庞大, 医学信息呈现多样化和非结构化的特点, 因此如何有效地从中文电子病历中提取医学信息, 成为医学信息处理和自然语言处理领域的重要研究方向之一.

命名实体识别 (named entity recognition, NER) 是从文本中识别并提取出具有特定意义的实体名称的任务, 是信息抽取、文本挖掘和自然语言处理中的一项基本任务. 电子病历命名实体识别旨在从电子病历中自动化地识别和提取出病人信息、疾病名称、药品名称、手术名称等医学实体信息. 这对于医疗机构、医生和研究人员来说都具有重要意义, 不仅可以协助医生进行临床分析来提高医疗诊断效率, 而且能够加快智慧医疗研究的进展.

目前, 许多研究者已经对电子病历命名实体识别进行了大量的研究, 并提出了各种方法和算法. 其中, 深度学习算法 (如卷积神经网络、循环神经网络和注意力机制等) 的出现, 使得电子病历命名实体识别的准确率和效率得到了显著提高. 但是, 中文电子病历命名实体识别领域还存在一些挑战和难点, 如医疗嵌套实体难以处理、中文实体边界的模糊性和实体类别的不平衡性等.

因此, 针对中文电子病历中医疗嵌套实体难以处理的问题, 本文基于 RoBERTa-wwm-ext-large 预训练模型提出一种知识增强的中文电子病历命名实体识别模型 ERBEGP. ERBEGP 借助于 EGP 从全局的角度出发, 同时考虑实体的起始和终止位置, 利用头部和尾部的特征信息来预测嵌套实体. 同时结合知识图谱和 BiLSTM 以提高模型的准确性和高效性. 在 cMedQANER、cEHRNER、cMeEE 和 cMeEE-V2 这 4 个数据集上进行实验验证, $F1$ 值分别达到了 81.22%、80.97%、67.03% 和 67.28%.

2 相关工作

2.1 预训练模型

预训练模型 (pre-trained model) 是指使用大规模语料库先对模型进行训练, 再通过迁移学习的方法将模型应用于特定任务中的一种模型. 目的是学习语言的潜在结构和规律, 并从中抽象出通用的语言表示形式, 以便在各种自然语言处理任务中使用. 预训练模型可以使得自然语言处理在不同的任务上表现更加出色, 而无需进行特定任务的训练, 减少了人工标注数据的需求, 大大提高了效率. Google AI[1] 于 2018 年提出的一种预训练语言模型 BERT (bidirectional encoder representations from Transformers). 采用双向 Transformer 编码器进行预训练, 通过 MLM (masked language model) 和 NSP (next sentence prediction) 两个任务对大规模的文本数据进行无监督学习, 使得模型能够自动地学习到词汇、句子之间的语义关系和上下文信息. BERT 在多项自然语言处理任务上都取得了非常好的效果, 成为目前最先进的语言模型之一. 目前, 主流的预训练模型有 BERT、RoBERTa 和 ALBERT 等.

由于医学表述通常具有专业性、复杂性和多样性等特点, 这对预训练模型的算法设计和数据处理都提出了挑战. 李正民等人[2] 基于 BERT 提出的多特征融合模型 BERT-BiLSTM-IDCNN-Attention-CRF, 通过融合模型 BiLSTM 和迭代膨胀卷积 (IDCNN) 有效地获得了电子病历上下文特征和局部特征, 在 CCKS2020 数据集上相较于 BiLSTM-CRF 等基准模型 $F1$ 值提升 1.27%. 赵奎等人[3] 提出了一种改进的 BiLSTM-CRF 深度学习模型用于电子病历命名实体识别, 通过 BiLSTM 对病历文本进行特征提取和 CRF 对病历文本进行约束, 提高了电子病历命名实体识别有效性, 相对于传统的 BiLSTM-CRF 模型, 该模型在实体类别上的 $F1$ 值提升了 3%–11%. 张芳丛等人[4] 基于 RoBERTa 提出深度学习模型 RoBERTa-WWM-BiLSTM-CRF, 通过融合模型 RoBERTa-WWM、BiLSTM 和 CRF 有效解决了中文电子病历命名实体识别中存在的一词多义和词识别不全的问题.

2.2 知识图谱

预训练模型通常是通过大规模的文本语料库进行训练的, 例如 BERT 系列模型. 这些模型学习到了大量的语言知识, 包括词汇、语法和语义. 然而, 这些模型并不总是具有完整的语义知识, 比如一些专业领域的实体和关系等知识. 这就需要知识图谱来填补这些空白. 知识图谱是一种结构化的知识表示形式, 它以图形的形式呈现实体、概念和它们之间的关系. 知识图谱可以提供大量的结构化知识, 例如实体、关系、属性等等, 这些知识可以被用来扩展预训练模型的语义表示能力. 随着各种基于知识图谱的预训练模型相继出现, 医疗知识图谱也得到了快速发展.

2020 年, Lee 等人[5] 在 BERT 模型的基础上进行改进和微调, 结合医疗知识图谱提出专门用于处理生

物医学领域的文本数据的医疗预训练模型 BioBERT. BioBERT 的知识图谱数据主要是从 PubMed 和 PMC 等生物医学文献数据库中提取的, 这些数据集包含了大量的生物医学专业术语和实体, 可以更好地适应生物医学领域的特点. BioBERT 是第 1 个基于生物医学语料库的预训练模型[6]. 同年, 阿里巴巴 Zhang 等人[7]提出针对中文医学文本的预训练模型 MC-BERT, 同样以 BERT 模型为基础模型, 通过知识图谱同时将生物医学语料和生物医药实体知识注入到模型中进行训练. 在命名实体任务 cEHRNER 和 cMedQANER 上, 相较于其他经典模型识别效果得到了显著提高, 平均 *F*1 值达到 90%. 2021 年, Rasmy 等人[8]提出了生物医学领域的预训练模型 Med-BERT, 相比于 BioBERT, Med-BERT 的知识图谱数据也是采用 PubMed 和 PMC 等生物医学文献数据库. 但 Med-BERT 是直接采用医学领域语料进行模型预训练. 随后, 杨飞洪等人[9]采用 Med-BERT 模型进行中文电子病历命名实体识别, 在 cMedQANER 数据集上 *F*1 值达到 82.29%, 且实验结果表明模型对“药物”实体的识别率较高.

3 ERBEGP 命名实体识别模型

本文通过结合知识增强的 RoBERTa-wwm-ext-large[10]、BiLSTM[11] 和 EGP[6], 构建了一个电子病历命名实体识别模型 ERBEGP, 如图 1 所示.

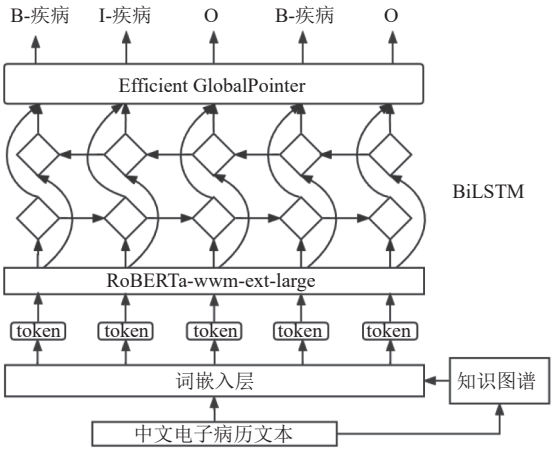


图 1 ERBEGP 模型总体架构

ERBEGP 模型中文电子病历命名实体识别基本流程如下.

1) 首先, 中文电子病历文本输入到模型后需要进行数据拷贝, 一部分作为初始电子病历文本直接送入词嵌入层, 另一部分交给知识图谱.

2) 知识图谱通过检索匹配相应的知识三元组, 将知识三元组与初始电子病历文本融合得到融合文本序列送入嵌入层, 以此获得数据集以外的医疗知识.

3) 经过 RoBERTa-wwm-ext-large 将输入的文本序列映射成高维度的向量表示, 通过全词掩码策略捕获文本中字级别和词级别的语义特征作为后续模型的输入.

4) RoBERTa-wwm-ext-large 的输出结果被输入到 BiLSTM 中进行序列编码, 通过 BiLSTM 的 3 种门控机制, 对电子病历文本选择性地进行遗忘或传递, 以此来捕获长序列文本依赖的特征信息, 能够更全面地理解文本中的上下文语义.

5) 最后经过 EGP 模型, 利用实体起始和终止位置的特征信息来预测嵌套实体, 更加灵活地进行命名实体识别任务中的序列标注, 最终得到标注序列.

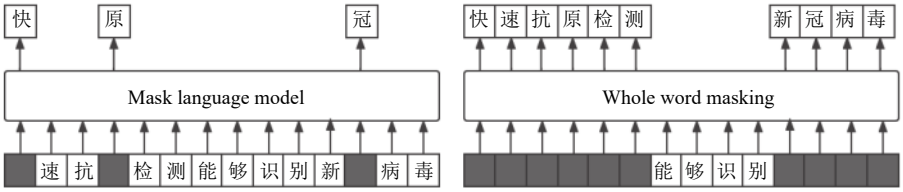
3.1 RoBERTa-wwm-ext-large 预模型

RoBERTa-wwm-ext-large[10] 是基于 RoBERTa 模型的中文预训练语言模型, 其名称中的“wwm”代表“whole word masking”, 即采用了全词掩盖策略进行训练. 相比于原始的 RoBERTa 模型, RoBERTa-wwm-ext-large 使用更大规模的语料库进行无监督训练, 同时采用了数据增强、全词掩盖策略以及动态掩盖策略等训练技巧, 以最大化语言模型对输入文本的理解. RoBERTa-wwm-ext-large 在多项中文自然语言处理任务上取得了优秀表现, 在命名实体识别任务上超越了 BERT、RoBERTa 等其他先进的预训练模型.

全词掩码策略可以缓解中文电子病历中信息丢失的问题. 在 BERT 的掩码策略 MLM 中[1], 将输入的句子中一些单词进行随机掩盖, 替换成特殊的掩码符号 Mask, 并要求模型预测这些单词, 从而使模型学习到对上下文的理解. 但是在中文任务中 MLM 随机掩码的是某一个字, 但在中文里有实际含义更有可能是这个字所组成的词或短语, 这样掩码策略可能会导致一些重要的信息被丢失. 而在 RoBERTa-wwm-ext-large 的全词掩码策略中, 通过使用分词工具 LTP[12] 识别词汇边界进行中文分词, 然后根据全词掩码策略将整个作为掩码单元进行掩码替换再进行预测. 全词掩码策略通过保留整个词组或短语的完整性而不仅仅是单个汉字特征, 缓解了中文电子病历中信息丢失的问题, 可以使模型更好地理解文本的上下文.

例如图 2 所示, 在模型的输入层输入病历文本“快速抗原检测能够识别新冠病毒”; 图 2(a) 中 BERT 利用 Wordpiece 进行分割, MLM 将“快”“原”和“冠”这

3 个字进行 Mask, 因此 BERT 在训练时学习到的更多是字与之间的关系. 而在图 2(b) 中 RoBERTa-wwm-ext-large 利用 LTP 进行中文分词, 然后通过全词掩码策略随机选取医疗实体名词“快速抗原检测”和“新冠病毒”进行 Mask, 因此 RoBERTa-wwm-ext-large 在训练时学习到的更多是词组与之间的关系.



(a) MLM 掩码策略示例图 (b) 全词掩码策略示例图

图 2 不同掩码策略对比图

3.2 ER 预训练语言模型

尽管 RoBERTa-wwm-ext-large 在自然语言处理任务中表现出色, 但是仍然存在一些改进的方向. 虽然 RoBERTa-wwm-ext-large 使用了大规模的无标注数据进行训练, 但是在某些特定任务和领域中, 如本文中的医疗领域, 可能需要更专业、更丰富的训练数据. 因此, 本文提出一个知识增强的预训练语言模型 ER (enhanced-RoBERTa-wwm-ext-large), 通过 RoBERTa-wwm-ext-large 与外部知识库结合起来, 利用外部知识更有效地提高模型的性能.

知识图谱可以看作是一种结构化的语义知识库, 通常以知识三元组 $(subject, predicate, object)$ 的形式进行表示, 其中 $subject$ 和 $object$ 是实体, $predicate$ 是它们之间的关系. 预训练模型引入知识图谱的策略可以分为预训练阶段引入和微调阶段引入[13]. 相比于预训练阶段引入, 在微调阶段引入知识图谱可以快速更换不同领域相关的知识库, 在各专业领域的下游任务中获得性能提升. 本文采用微调阶段引入知识图谱方法将知识三元组与 RoBERTa-wwm-ext-large 进行知识融合, 如图 3 所示. 基本方法如下.

1) 在知识检索层对初始语句 $s=\{c_0,c_1,c_2,\cdots,c_n\}$ 使用分词工具 LTP 识别词汇边界进行中文分词, 得到实体名词集合 $\mathbb{E}=\{Noun_0,Noun_1,Noun_2,\cdots,Noun_n\}$.

2) 对集合 $\mathbb{E}$ 中每个实体 $Noun_i$ 在知识图谱 $\mathbb{K}$ 中进行检索匹配 $E=K\_Query(\mathbb{E},\mathbb{K})$[14], 生成相应的三元组 $E=\{(w_i,r_{i0},w_{i0}),\cdots,(w_i,r_{ik},w_{ik})\}$.

3) 将检索到的知识三元组 E 通过绝对位置与相对位置与初始语句 s 进行融合, 生成句子树 t, $t=K\_Inject(s,E)$[14]. 绝对位置是指融合知识三元组后将句子按句子语序依次标记的位置. 相对位置是指保留三元组的字符顺序标记, 在初始语句之前的位置标记上直接进行拼接的位置.

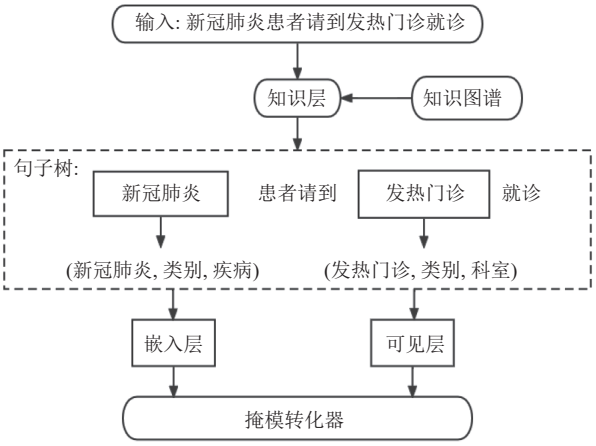


图 3 ER 模型总体架构

4) 将句子树送入嵌入层和可见层. 嵌入层的作用是保留结构化语义信息, 是融合文本、绝对位置和相对位置 3 个部分之和. 如图 4 所示, 其中, 下划线标记为知识三元组的字符顺序. 可见层的作用是通过构建可见矩阵防止知识噪声, 注入的知识过多, 但知识之间并无关联就会产生知识噪声, 造成句意改变. 例如在图 5 中融合文本“新冠肺炎类别疾病患者请到发热门诊类别就诊”中, “类别疾病”和“请到”之间并无关联, 因此需要屏蔽两者之间的注意力. 图中黑色表示横纵方向的两个字符可以进行注意力计算, 白色表示屏蔽两者注意力的计算.

5) 最后采用掩模转换器编码作为后续任务的输入, 掩模转换器是多层 Transformer 堆叠而成, 用于步骤 4) 可见矩阵注意力的计算, 如式 (1), 式 (2)[14] 所示. 其中 $h^l\in R^{L\times d_m}$ 是第 l 层 Transformer 的输出. Q_i^{l+1},

K_i^{l+1}和V_i^{l+1}代表第$l+1$层的第i个注意力头的$Query$矩阵, Key矩阵和$Value$矩阵, $W_i^{Q,l+1}$, $W_i^{K,l+1}$和$W_i^{V,l+1}$表示与之对应的转换矩阵作为模型参数. 计算在第$l+1$层中的每个头的注意力时, 将上述参数送入第$l+1$层中, 通过添加可见矩阵$M \in R^{L\times L}$来达到控制遮盖注意力, L为序列长度.

$$\begin{cases} Q_i^{l+1} = h^l W_i^{Q,l+1} \\ K_i^{l+1} = h^l W_i^{K,l+1} \\ V_i^{l+1} = h^l W_i^{V,l+1} \end{cases} \tag{1}$$

$$A_i^{l+1} = Softmax\left(\frac{Q_i^{l+1}(K_i^{l+1}) + M}{\sqrt{d_k}}\right)V_i^{l+1} \tag{2}$$

3.3 双向长短时记忆网络 (BiLSTM) 模型

针对循环神经网络 (RNN) 无法处理训练过程中产生梯度消失或梯度爆炸的问题, 1997 年 Hochreiter 等人提出长短时记忆网络 (long short-term memory, LSTM)[11], 该模型是对 RNN 缺陷的改进, 通过引入输入门、输出门和遗忘门 3 种门控机制, 对输入的语义信息选择性地进行遗忘或传递, 以此来捕获长序列文本依赖的特征信息, 能够更全面地理解序列的语义.

绝对位置:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
融合文本:	新	冠	肺	炎	类	别	疾	病	患	者	请	到	发	热	门	诊	类	别	科	室	就	诊
相对位置:	1	2	3	4	**5**	**6**	**7**	**8**	5	6	7	8	9	10	11	12	**9**	**10**	**11**	**12**	13	14

图 4 嵌入层位置信息融合示例图

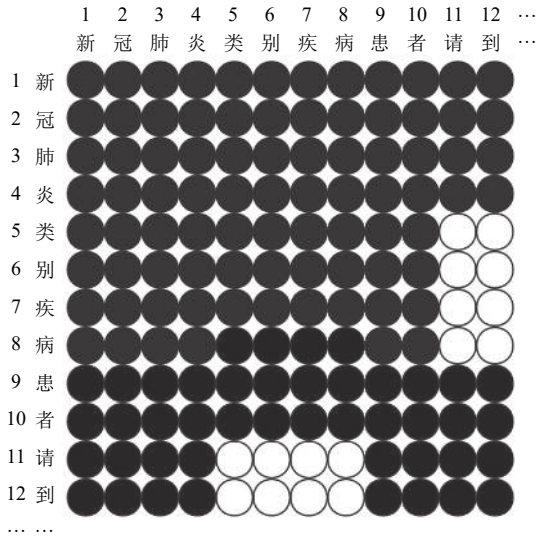


图 5 可见矩阵示例图

LSTM 虽然缓解了梯度消失或梯度爆炸的问题, 但是 LSTM 只能从左向右或从右向左处理输入序列, 不能同时考虑电子病历中前后文信息, 因此无法捕获到电子病历中命名实体的一些关键特征. BiLSTM 基于 LSTM 原理由两个相互独立的 LSTM 组成, 一个从左向右正向处理输入序列, 另一个从右向左反向处理输入序列, 最终得到一个包含电子病历上下文信息的综合语义特征.

3.4 全局指针网络 (EGP) 模型

嵌套命名实体识别是中文电子病历命名实体识别中的较为棘手的一个子任务. 嵌套命名实体是一种特殊的命名实体, 即在一个实体的内部至少还存在着一个其他的实体. 例如“呼吸中枢受累”的实体类型是症状, 其中还包含着类型为部位的命名实体“呼吸中枢”. 针对这一问题, Su 等人[6] 提出全局指针网络模型 (global pointer, GP). GP 可以有效地识别嵌套实体和非嵌套实体, 且对于非嵌套实体 (Flat NER) 的识别效果可以媲美 CRF[15]. GP 从全局的角度出发, 同时考虑实体的起始和终止位置, 利用头部和尾部的特征信息来预测嵌套实体. 由于 GP 的训练和预测过程都是并行的, 无需像 CRF 一样进行递归运算, 因此更加高效. GP 新增实体类型时模型的参数量也会随着增加, 针对原 GP 参数利用率不高的问题, 在 GP 基础进行改进提出 EGP, 明显降低了 GP 的参数量. 多个命名实体识别任务的实验结果表示, 参数量更少的 EGP 反而还取得更好的效果.

对于不同类别的实体, GP 通过不同的 Head 进行预测, 如图 6 所示. 其中, “呼吸中枢受累”的实体类型为症状, “呼吸中枢”的实体类型为部位. 定义$s_\alpha(u,v)$是一个实体类型为α, 字符序列从u到v的打分. 其中u为图中横坐标, v为图 6 中纵坐标, 为 1 表示该坐标下$s(u,v)$是α类型的实体, 为 0 则不是. GP 的具体实现如式 (3)–式 (6) 所示[12].

$$h_1, h_2, \cdots, h_n = PLM(x_1, x_2, \cdots, x_n) \tag{3}$$

$$q_{u,\alpha} = W_{q,\alpha}h_u + b_{q,\alpha} \tag{4}$$

$$k_{v,\alpha} = W_{k,\alpha}h_v + b_{k,\alpha} \tag{5}$$

$$s_\alpha(u,v)=(\mathfrak{R}_u q_{u,\alpha})^{\mathrm{T}}(\mathfrak{R}_v k_{v,\alpha})=q_{u,\alpha}^{\mathrm{T}}\mathfrak{R}_{v-u}k_{v,\alpha} \tag{6}$$

其中, $PLM(x_1,x_2,\cdots,x_n)$表示经过预训练模型处理后的电子病历输入,得到一个长度为 n 的向量序列$h_1,h_2,\cdots,h_n$. 对于每一个h_i会生成起始位置表征信息$q_{u,\alpha}$和终止位置表征信息$k_{v,\alpha}$, $W_{q,\alpha}$和$W_{k,\alpha}$是变换矩阵, 通过对$q_{u,\alpha}$和$k_{v,\alpha}$进行内积得到实体类型α为字符序列从u到v的得分. 此外为了充分利用电子病历命名实体的边界信息, 在实体得分中引入旋转位置编码 (RoPE)[12], 利用变换矩阵$\mathfrak{R}_u$的转秩矩阵和$\mathfrak{R}_v$相乘得到$\mathfrak{R}_{v-u}$, 将其分别应用到$q_{u,\alpha}$和$k_{v,\alpha}$中,得到包含u到v的相对位置信息的得分$s_\alpha(u,v)$.

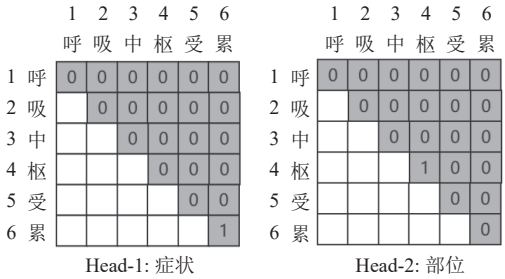


图 6　EGP 实体预测图

由于 GP 每增加一种实体类型就需要得到两个变换矩阵, 当类型较多时, 相应的参数量就会增多. 为了解决这一问题, 提出了矩阵共享的方法. 即当对于同一实体类型α进行识别时, 通过共用一个打分矩阵$(W_q h_u)^{\mathrm{T}}(W_k h_v)$, 该矩阵仅用于实体预测, 即使新增实体类型也可以继续沿用, 加上一个针对类型α的偏置矩阵实现类型预测, 以此实现参数量的缩减. 为了进一步缩减参数量, 使用$[q_u;k_u]$来代替h_u. 如式 (7), 式 (8) 所示[6].

$$s_\alpha(u,v)=(W_q h_u)^{\mathrm{T}}(W_k h_v)+w_\alpha[h_u;h_v] \tag{7}$$

$$s_\alpha(u,v)=q_u^{\mathrm{T}}k_v+w_\alpha^{\mathrm{T}}[q_u;k_u;q_v;k_v] \tag{8}$$

4　实验与结果

4.1　数据集

本文实验数据均来自于中文医疗信息处理评测基准 CBLUE (Chinese biomedical language understanding evaluation)[16]. 本文选取了 CBLUE 中的 cMedQANER、cEHRNER、cMeEE 和 cMeEE-V2 这 4 个数据集进行验证. 数据集的训练集、验证集、测试集和评价标准如表 1 所示.

由于 4 个数据集格式不同, 本文实验将数据集进行数据预处理, 将数据集全部统一转换成 BIO 标注格式. BIO 标注格式是一种常见的文本标注方式[6], 主要用于命名实体识别任务. BIO 是基于词的标注方式, 它将每个词进行标注, 用于表示该词是否属于某个命名实体. B-命名实体类型: 表示当前词是一个命名实体的开始. I-命名实体类型: 表示当前词是一个命名实体的中间部分. O: 表示当前词不是命名实体的一部分.

表 1　数据集相关参数

数据集	训练集	验证集	测试集	评价标准
cMedQANER	1 673	175	175	*F*1
cEHRNER	914	44	41	*F*1
cMeEE	15 000	5 000	3 000	*F*1
cMeEE-V2	15 000	5 000	3 000	*F*1

4.2　实验设置

实验采用 RoBERTa-wwm-ext-large 模型对数据集进行预训练处理, 最终得到 512 维的向量表示. 在 BiLSTM 层采用 Dropout 正则化减少过度拟合, 设置 LSTM 隐层单元数为 512, Dropout 为 0.1. 实验采用 Adam 作为优化器, 学习率为 1E–5, 批处理大小 batch_size 为 32, 训练轮次 epochs 为 10.

4.3　评测指标

实验采用 *F*1 值作为模型识别效果的评价标准, 如式 (9)–式 (11) 所示:

$$P=\frac{T_P}{T_P+F_P}\times 100\% \tag{9}$$

$$R=\frac{T_P}{T_P+F_N}\times 100\% \tag{10}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{11}$$

其中, T_P表示验证集中被正确预测为正类的实体个数; F_P表示验证集被负类错误预测为正类的实体个数; F_N表示验证集被正类错误预测为负类的实体个数. P表示准确率; R表示召回率; $F1$值为P, R的调和平均数, 兼顾了准确率和召回率.

4.4　实验结果及分析

由于数据集的训练集、验证集和测试集划分比例不同, 验证集和测试集比例较小的 cMedQANER、cEHRNER 的准确率明显高于 cMeEE 和 cMeEE-V2. 相较于基础模型 RoBERTa-wwm-ext-large, 本文基于其提出的 ERBEGP 模型在 4 个数据集上 *F*1 值分别提高了 2.28%、5.25%、5.20% 和 4.74%. 相较于经典 BERT, ERBEGP 在 4 个数据集上 *F*1 值分别提高了 6.21%、5.55%、4.92% 和 5.27%. 其结果如表 2 所示.

在消融实验中可以看出结合了医疗知识图谱的

RoBERTa-wwm-ext-large 模型性能得到明显的提升. 相较于基础模型 RoBERTa-wwm-ext-large, ER 模型在 4 个数据集上的 *F*1 值均得到了较大的提升, 尤其是在 cEHRNER 数据集上 *F*1 值提升了 1.95%. 由于 RoBERTa-wwm-ext-large 采用无监督训练, 通过在微调阶段引入知识图谱, 利用类似于远监督的方式对 RoBERTa-wwm-ext-large 进行有监督训练, 从而学习到了更多医疗实体, 进一步提高了模型的性能.

表 2 实验结果 (%)

实验	模型	cMedQANER	cEHRNER	cMeEE	cMeEE-V2
消融实验	RoBERTa-wwm-ext-large	78.94	75.72	61.83	62.54
	ER	79.92	77.67	63.42	64.48
	ER+BiLSTM	80.43	78.98	64.83	65.26
	ER+BiLSTM+CRF	80.62	80.57	65.95	66.42
对比实验	BERT	75.01	75.42	62.11	62.01
	ALBERT-xlarge	64.62	62.52	61.83	62.13
	RoBERTa	76.48	76.55	62.11	62.02
	RoBERTa-wwm-ext-base	77.72	77.42	62.42	62.33
本文实验	ERBEGP	81.22	80.97	67.03	67.28

BiLSTM 对模型的提升虽然不如知识图谱显著, 但也使得模型识别效果有了较大提高. 采用 BiLSTM 与 ER 结合的方式, BiLSTM 能够更好地捕获句子中的上下文信息, 缓解了识别过程中梯度消失或梯度爆炸的问题, 进一步增强了模型对文本语义的理解.

相较于 CRF, EGP 对中文电子病历命名实体识别的效果明显更好. 在同等条件下, ERBEGP 比 ER+BiLSTM+CRF 在 4 个数据集上的 *F*1 值提高了 0.6%、0.4%、1.08% 和 0.86%. 实验结果表明在中文电子病历中包含了大量医疗嵌套实体, 而 EGP 能够同时处理医疗实体中的嵌套实体和非嵌套实体, 且其训练和预测过程都是并行的[6], 因此 EGP 在中文电子病历命名实体识别任务上的更加准确和高效.

大模型的识别效果通常比小模型的识别效果好. 在对比实验中, 相较于 BERT, ALBERT-xlarge[17] 采用了更小的预训练数据集, 更少的参数量和更短的训练时间, 但是实验效果并不理想, 在 4 个数据集上 ALBERT-xlarge 的 *F*1 值基本均低于 BERT. 而 RoBERTa 使用了更大的预训练数据集, 更长的训练时间, 进行了更多的预训练轮次, 在本文所有数据集上的识别效果均高于 BERT. 实验证明, 越多预训练数据集电子病历识别效果更好, 所以本文采用了比 RoBERTa 更大的预训练模型 RoBERTa-wwm-ext-large.

采用全词掩码策略的模型更适用于中文自然语言处理任务. RoBERTa-wwm-ext-base 相比于 RoBERTa 在 4 个数据集上的 *F*1 值提高了 1.24%、0.87%、0.31% 和 0.31%. 由于 RoBERTa-wwm-ext-base 使用了全词掩码策略, 获得了词级别的向量表示, 保留汉字中整个词组或短语的完整特征, 减轻了 RoBERTa 中只能获得字符级向量的缺点, 缓解中文电子病历中信息丢失的问题, 因此更适合应用于中文医疗领域命名实体识别任务.

实验结果表明 ERBEGP 模型相较于经典模型 BERT、ALBERT-xlarge、RoBERTa 和 RoBERTa-wwm-ext-base, 识别效果均有显著提升, 在 cMedQANER、cEHRNER、cMeEE 和 cMeEE-V2 这 4 个数据集上 *F*1 值分别为 81.22%、80.97%、67.03% 和 67.28%, 均达到了最优.

对比实验中不同模型在 cMedQANER 数据集的实验结果如图 7 示. 各个对比模型中 ALBERT-xlarge 的参数量为 59M, 在第 4 轮训练时 *F*1 达到最大值 64.62% 并收敛. ERBEGP 的参数量为 325M, 在第 6 轮训练时收敛, 第 7 轮训练时 *F*1 达到最大值 81.22%. 由于 ALBERT-xlarge 参数量最小, 更容易学习数据集中的规则与特征[17], 达所以收敛的速度最快. ERBEGP 虽然模型参数量大且复杂程度高, 但是由于引入了知识图谱, 使得模型率先学到了更多的医疗知识, 这些先验知识使得 ERBEGP 具有更好的初始状态, 从而导致模型第 1 轮训练时具有较高的 *F*1 值, 由于 ERBEGP 需要优化大量参数, 因此需要迭代更多轮次, 收敛速度较慢.

5 结束语

本文基于 RoBERTa-wwm-ext-large 提出一种知识增强的实体识别模型 ERBEGP, 用于中文电子病历命名实体识别. RoBERTa-wwm-ext-large 的全词掩码策略更加容易捕获中文词或短语的语义特征, 缓解了中文电子病历中信息丢失的问题. 通过结合知识图谱使模型学习到了更多的医疗知识, 进一步提高模型对电

子病历实体识别的准确性; 利用 BiLSTM 的 3 种门控机制, 能够更好捕获病历综合电子病历语义特征; 采用 EGP 同时考虑实体的起始和终止位置, 利用头部和尾部的特征信息来预测嵌套实体, 更加有效地解决中文电子病历命名实体识别任务中嵌套实体难以处理的问题, 使模型在中文电子病历命名实体识别任务上的识别效果更加准确和高效. 在 CBLUE 中的 4 个数据集上本文方法均取得了更好的识别效果, 对中文电子病历的嵌套实体识别具有一定的参考价值.

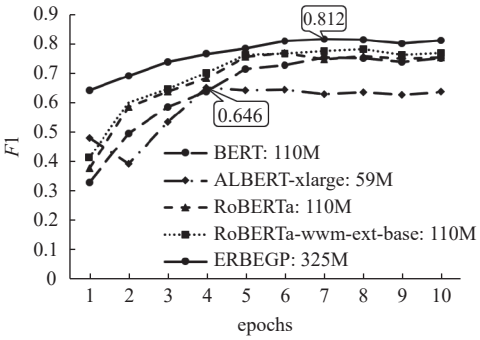


图 7 cMedQANER 对比实验结果

随着以 ChatGPT 等为代表的大规模预训练模型的出现, 逐渐掀起了预训练模型朝大规模方向发展的浪潮. 大量研究表明, 模型参数量越大, 训练数据量越多的预训练模型表现更出色. 如何训练这些大模型用于中文电子病历命名实体识别是接下来的研究方向.

参考文献

1 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186.

2 李正民, 云红艳, 王翊臻. 基于 BERT 的多特征融合的医疗命名实体识别. 青岛大学学报 (自然科学版), 2021, 34(4): 23–29. [doi: 10.3969/j.issn.1006-1037.2021.11.05]

3 赵奎, 杜昕娉, 高延军, 等. 融合文字与标签的电子病历命名实体识别. 计算机系统应用, 2022, 31(10): 375–381. [doi: 10.15888/j.cnki.csa.008723]

4 张芳丛, 秦秋莉, 姜勇, 等. 基于 RoBERTa-WWM-BiLSTM-CRF 的中文电子病历命名实体识别研究. 数据分析与知识发现, 2022, 6(2–3): 251–262. [doi: 10.11925/infotech.2096-3467.2021.0910]

5 Lee J, Yoon W, Kim S, *et al*. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020, 36(4): 1234–1240. [doi: 10.1093/bioinformatics/btz682]

6 Su JL, Murtadha A, Pan SF, *et al*. Global pointer: Novel efficient span-based approach for named entity recognition. arXiv:2208.03054, 2022.

7 Zhang NY, Jia QH, Yin KP, *et al*. Conceptualized representation learning for Chinese biomedical text mining. arXiv:2008.10813, 2020.

8 Rasmy L, Xiang Y, Xie ZQ, *et al*. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine, 2021, 4(1): 86. [doi: 10.1038/s41746-021-00455-y]

9 杨飞洪. 面向中文临床自然语言处理的 BERT 模型研究 [硕士学位论文]. 北京: 北京协和医学院, 2021. [doi: 10.27648/d.cnki.gzxhu.2021.000896]

10 Cui YM, Che WX, Liu T, *et al*. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504–3514. [doi: 10.1109/TASLP.2021.3124365]

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

12 Che WX, Li ZH, Liu T. LTP: A Chinese language technology platform. Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Beijing: ACL, 2010. 13–16.

13 Zhang W, Wong CM, Ye GQ, *et al*. Billion-scale pre-trained e-commerce product knowledge graph model. Proceedings of the 37th IEEE International Conference on Data Engineering. Chania: IEEE, 2021. 2476–2487.

14 Liu WJ, Zhou P, Zhao Z, *et al*. K-BERT: Enabling language representation with knowledge graph. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020. 2901–2908.

15 Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning. Williamstown, 2001. 282–289.

16 Zhang NY, Chen MS, Bi Z, *et al*. CBLUE: A Chinese biomedical language understanding evaluation benchmark. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022. 7888–7915.

17 Lan ZZ, Chen MD, Goodman S, *et al*. ALBERT: A lite BERT for self-supervised learning of language representations. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.

(校对责编: 孙君艳)