计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(11):108−119 [doi: 10.15888/j.cnki.csa.009285]

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

基于 MFE-BERT 与 FNNAttention 的心理医学知识图谱构建①

刘子轩[1], 申艳光[1], 李　焰[2], 苏文婷[1,3]

[1](河北工程大学 信息与电气工程学院, 邯郸 056038)
[2](河北工程大学附属医院 急诊科, 邯郸 056038)
[3](河北工程大学 河北省安防信息感知与处理重点实验室, 邯郸 056038)
通信作者: 申艳光, E-mail: 598991558@qq.com

摘　要: 针对心理医学领域文本段落冗长、数据稀疏、知识散乱且规范性差的问题, 提出一种基于多层级特征抽取能力预训练模型 (MFE-BERT) 与前向神经网络注意力机制 (FNNAttention) 的心理医学知识图谱构建方法. MFE-BERT 在 BERT 模型基础上将其内部所有 Encoder 层特征进行合并输出, 以获取包含更多语义的特征向量, 同时对两复合模型采用 FNNAttention 机制强化词级关系, 解决长文本段落语义稀释问题. 在自建的心理医学数据集中, 设计 MFE-BERT-BiLSTM-FNNAttention-CRF 和 MFE-BERT-CNN-FNNAttention 复合神经网络模型分别进行心理医学实体识别和实体关系抽取, 实体识别 $F1$ 值达到 93.91%, 实体关系抽精确率达到了 89.29%, 通过融合文本相似度与语义相似度方法进行实体对齐, 将所整理的数据存储在 Neo4j 图数据库中, 构建出一个含有 3 652 个实体, 2 396 条关系的心理医学知识图谱. 实验结果表明, 在 MFE-BERT 模型与 FNNAttention 机制的基础上构建心理医学知识图谱切实可行, 提出的改进模型所搭建的心理医学知识图谱可以更好地应用于心理医学信息管理中, 为心理医学数据分析提供参考.

关键词: 知识图谱; 心理医学; 命名实体识别; 实体关系抽取; MFE-BERT 模型; FNNAttention 机制

引用格式: 刘子轩,申艳光,李焰,苏文婷.基于 MFE-BERT 与 FNNAttention 的心理医学知识图谱构建.计算机系统应用,2023,32(11):108–119. http://www.c-s-a.org.cn/1003-3254/9285.html

Construction of Psychomedical Knowledge Graph Based on MFE-BERT and FNNAttention

LIU Zi-Xuan[1], SHEN Yan-Guang[1], LI Yan[2], SU Wen-Ting[1,3]

[1](School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China)
[2](Department of Emergency, Affiliated Hospital of Hebei Engineering University, Handan 056038, China)
[3](Hebei Key Laboratory of Security and Protection Information Sensing and Processing, Hebei University of Engineering, Handan 056038, China)

Abstract: To solve the problems of lengthy paragraphs, sparse data, scattered knowledge, and poor specification of text data in psychological medicine, a method based on the pre-trained model of multi-level feature extraction capability (MFE-BERT) and forward neural network attention (FNNAttention) mechanism is proposed for the construction of psychomedical knowledge graphs. Based on the BERT model, MFE-BERT merges and outputs all the internal encoder layer features to obtain feature vectors with more semantics. At the same time, the FNNAttention mechanism is applied to the two composite models to strengthen the word-level relationship and solve the semantic dilution of long text paragraphs. In the self-created psychomedical datasets, the compound neural network models of MFE-BERT-BiLSTM-FNNAttention-CRF and MFE-BERT-CNN-FNNAttention are designed for psychomedical entity recognition and entity

① 基金项目: 国家自然科学基金 (61802107); 河北省医学科学研究课题 (20220037); 国家重点研发计划 (2018YFF0301004)
收稿时间: 2023-04-13; 修改时间: 2023-05-17; 采用时间: 2023-05-25; csa 在线出版时间: 2023-09-19
CNKI 网络首发时间: 2023-10-07

relationship extraction respectively. The entity recognition *F*1 value reaches 93.91% and the entity relation extraction precision rate reaches 89.29%. The entity alignment is carried out by merging text similarity and semantic similarity. The collated data are stored in a Neo4j graph database, and a psychomedical knowledge graph containing 3 652 entities and 2 396 relationships is constructed. The experimental results show that it is practical and feasible to construct a psychomedical knowledge graph based on the MFE-BERT model and the FNNAttention mechanism, and the psychomedical knowledge graph built by the proposed improved models can be better applied in psychomedical information management, providing a reference for psychomedical data analysis.

Key words: knowledge graph; psychomedical; named entity recognition; entity relationship extraction; MFE-BERT model; forward neural network attention (FNNAttention) mechanism

随着医学水平的不断提高以及互联网科学技术的高速发展, 心理学不仅在医学领域成为重点研究学科, 在融合互联网、数据挖掘[1]、神经网络等新兴科技的心理与行为研究被更提升至前所未有的高度. 面对心理医学知识管理的复杂性以及领域信息收集与整理效率低下的现状, 心理科学研究和分析对心理医学数据提出更高的要求.

当前心理医学领域迫切需要一种高效便捷的信息管理体系. 拥有强大语义网络的知识图谱[2] 能够为复杂的心理医学数据建立网络神经关系框架, 更好地为心理医学研究提供数据关系支撑. 徐春等人[3] 通过融合 BERT-WWM 和指针网络的实体关系联合抽取模型构建了旅游知识图谱. Martnez-Rodrguez 等人[4] 使用开放信息提取 (OpenIE) 生成的二元关系方法构建了文本关系知识图谱. 黄梦醒等人[5] 利用 BiLSTM-CRF 的实体识别与关系抽取方法构建电子病例知识图谱, 为个性化医疗推荐服务提供帮助. 廖开际等人[6] 综合利用 BiLSTM、BiGRU、CRF 等深度学习模型对社区医疗文本进行实体识别与关系抽取, 通过图数据库构建成可视化知识图谱.

近年来, 由结构化数据组成的知识图谱逐渐被科研人员应用于生物医疗等垂直领域, 但是由于心理医学文本段落冗长、数据稀疏、知识散乱的特点, 常见深度学习模型难以学习到心理医学数据集全部特征, 无法支撑心理医学知识图谱构建工作, 因此目前心理医学领域方面的知识图谱研究少之又少, 无法满足当前社会对心理医学信息管理的需求.

为解决上述存在的问题, 本文将深度神经网络学习与心理医学实体识别、心理医学实体关系抽取相结合, 提出具有多层级特征抽取能力的 MFE-BERT 预训练模型和应用前向神经网络的注意力机制改进模型 FNNAttention 进行心理医学实体识别与心理医学实体关系抽取, 最终通过两模型中识别的实体及其关系形成的结构化数据搭建心理医学知识图谱. 相比于传统模型, 本文改进模型做出以下贡献.

(1) MFE-BERT 模型将文本预处理生成动态词向量, 使心理医学实体识别和心理医学实体关系抽取的词向量融合上下文语义联系. 提出多层级特征抽取的改进预训练模型, 在全连接 Transformer Encoder 每层特征信息的基础上, 将每层的输出向量进行最终合并输出, 赋予词向量更为丰富的词级与语义信息.

(2) 应用 FNNAttention 模型, 利用前向神经网络自适应学习函数分配特征权重, 可以有效解决长文本段落语义稀释问题, 避免相同心理医学实体在不同语句中标注不一致的问题. 本文对传统 Attention 机制进行改进, 采用前向神经网络注意力机制[7] 捕获全局词级信息来强化长文本上下文的词级关系.

(3) 基于自建的心理医学数据集以及公开的生物医学数据集, 将提出的模型与其他已有基准模型进行对比分析, 实验结果证明本文所提出模型的合理性及有效性.

1 相关工作

1.1 命名实体识别

命名实体识别是将文本中的命名实体定位并分类为预定义实体类别的过程[8]. 近年来研究专家对深度学习不断探索且取得了良好进展, 部分学者尝试着将深度学习与命名实体识别相结合. Collobert 等人[9] 使用 CNN 对输入序列进行特征提取, 再通过 CRF 随机条件场合输出序列的标签. 这种模型对局部信息特征提取效果较好, 但是可能会忽略重要的上下文信息, 更适用于简单的表面特征抽取应用. Habibi 等人[10] 在生物医

学领域中利用 BiLSTM-CRF 模型对医学实体进行识别, 相较于 CNN 模型, BiLSTM 能够有效地捕捉句子和段落之间的关系, 更好地利用上下文语境进行实体识别, 但是 BiLSTM 模型容易出现过拟合问题, 因此不适合过小的数据集. 郭知鑫等人[11] 通过 BERT-BiLSTM-CRF 模型有效对法律文本中的案件实体进行识别, 提高了案件处理的效率. 任媛等人[12] 在 BERT-BiLSTM-CRF 模型的基础上引入了注意力机制, 在渔业标准定量指标的实体识别上做出了贡献. 心理医学领域, Lakel 等人[13] 在心理科学词典的基础上, 利用 JAPE (Java 注释模式引擎) 规则来提取心理实体, 但是需要花费大量时间和精力来定义规则, 难以面对复杂的数据.

心理医学实体识别技术对于心理健康领域的临床和研究具有广泛的应用前景. 通过识别文本中的心理医学实体, 如疾病名称、治疗方法、药品等, 可以为医生提供辅助诊断的依据, 同时也可以从大量的文本数据中提取知识和信息, 帮助构建更全面和准确的心理健康知识库.

1.2 实体关系抽取

关系抽取技术是搭建知识图谱过程中的重要一步, 也是自然语言处理任务的支撑基础. 关系抽取将文本中结构化、半结构化和非结构化的数据信息转化为具有结构化关系的数据信息存储在知识库中, 为之后的智能检索和语义分析提供一定的支持和帮助[14]. 关系抽取具体定义如下: 对于一个非结构化文本语句 S, 句中包含给定的实体对 ($E1$, $E2$), 提取两实体之间的关系 $r \in R$ (R 为预定义的关系集合), 形成一个关系三元组<$E1$, r, $E2$>.

目前基于深度学习的关系抽取节省了大量的时间以及人力成本, 在提高准确率的同时其模型泛化性得到了很好的拓展[15]. 陆晓蕾等人[16] 在多层级专利分类研究中将 Word2Vec-CNN 与 BERT-CNN 做对比实验, 后者表现更好, 准确率达到 84.3%. Zhou 等人[17] 提出了融合注意力与 BiLSTM 模型来捕捉句子中最重要的语义信息, 在 SemEval-2010 分类任务中 $F1$ 值达到 82.5%. 刘峰等人[18] 在已有研究的基础上通过引入 Multi-head Attention 和依存句法特征, 能够获取更多的文本句法信息. 姚宁等人[19] 对 CNN 模型应用注意力机制, 在精神分裂症分类任务中拥有较高的分类精度, 为临床针对提供生物学依据.

心理医学领域实体关系抽取是一种基于自然语言处理技术的医学领域应用, 旨在对心理医学相关文本中的实体间的关系进行分类. 它可以帮助医生、研究人员和决策者更好地理解心理疾病的发生和发展机制, 提供更加准确、全面的支持.

2 心理医学知识图谱模型构建方法

心理医学知识图谱构建的主要过程包括实体抽取和实体间关系的建立[20]. 首先通过 MFE-BERT-BiLSTM-FNNAttention-CRF 模型对心理医学文本进行实体识别, 其次将识别的实体与文本输入至 MFE-BERT-CNN-FNNAttention 模型进行实体之间的关系抽取, 最后进行知识融合, 利用 Neo4j 图数据库存储数据, 构成心理医学知识图谱.

2.1 心理医学实体识别模型

心理医学实体识别模型主要由 4 层组成, 分别是 MFE-BERT 预处理层, 双向长短记忆神经网络层, FNN-Attention 层和 CRF 条件随机场, 如图 1 所示.

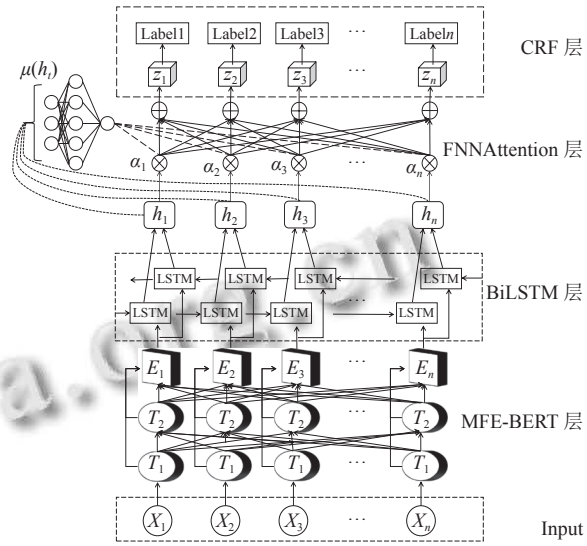


图 1 MFE-BERT-BiLSTM-FNNAttention-CRF 模型图

各层操作步骤如下.

(1) 文本输入层: 将心理医学文本以句子为单位作为输入序列.

(2) MFE-BERT 层: 接收文本输入层序列, 首先将序列向量化, 结合 Token Embeddings、Segment Embeddings 与 Position Embeddings 这 3 个向量作为输入, 经过对模型内部的改进, 每个向量经过 12 个 Transformer Encoder 层的输出被拼接成为一个具有当前序列深层特征语义的向量.

(3) BiLSTM 层: 接收 MFE-BERT 模型输出的特征向量, 通过双向长短记忆网络捕获心理医学文本的上下文语义表征, 输出每个实体标签的分数 (Emission_score).

(4) FNNAttention 机制: 采用前向神经网络注意力机制对 BiLSTM 层运算向量进行加权平均处理, 对长文本语义信息进行强化, 同时避免相同心理医学实体在不同语句中标注不一致的问题.

(5) CRF 层: 将经过权重分配处理的 Emission_score 作为输入, 通过对标签建模以及向最终的预测标签添加一些约束, 输出符合标注转移约束条件、最大可能的标注序列.

(6) 输出层: 实体识别模型整体能够将复杂的心理医学文本处理为最优标注序列, 最终根据标心提取医学实体.

2.2 心理医学关系抽取模型

心理医学关系模型由 3 部分组成, 分别是 MFE-BERT 预处理层, 卷积神经网络层 (CNN)[21] 和前向神经网络注意力机制. 关系抽取复合模型同样应用了上述改进的 MFE-BERT 预处理层和前向神经网络注意力机制来进行长文本语义信息的深度提取, 模型图如图 2 所示.

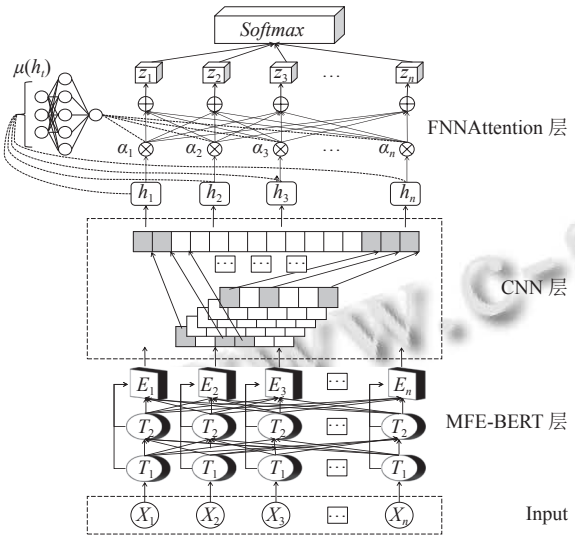


图 2 MFE-BERT-CNN-FNNAttention 模型图

各层操作步骤如下.

(1) 文本输入层: 利用心理医学实体识别模型处理以后的心理医学文本作为模型输入, 其结构为实体 1, 实体 2, 文本句子. 实体与实体, 实体与句子之间用“*”分隔, 句子中的心理医学实体用“#”表示, 如“忧郁症*中年人*###常见于###”这种方式使关系抽取模型在不进行额外距离标注的情形下也可以很好地学习到输入序列中两实体的关系.

(2) MFE-BERT 层: 文本传输给 MFE-BERT 模型中, 文字此时会被模型处理为具有词向量、位置向量以及句子向量结合而成的特征向量, 通过 12 层 Transformer Encoder 对特征向量进行预训练.

(3) CNN 层: 接收 MFE-BERT 模型输出的特征向量, 先利用卷积核对句子进行卷积操作以及池化层的降维和特征融合, 提取序列局部特征, 再通过多层卷积学习到文本的全局特征.

(4) FNNAttention 层: 采用前向神经网络注意力机制加强对心理医学关系分类作用明显的字的权重, 同时强化长文本的语义信息.

(5) 输出层: 采用 *Softmax* 函数作为分类器, 将 FNNAttention 层输出向量进行输入, 确定句中实体间特定的心理医学关系类型.

2.3 实体识别及关系抽取主要方法

在心理医学实体识别与心理医学关系两部分自然语言处理任务中都应用了 MFE-BERT 模型及 FNN-Attention 机制, 作为核心模型方法, 其对心理医学长文本数据的深度特征提取起着重要作用.

2.3.1 MFE-BERT 模型

对于心理医学文本, 其语句冗长且复杂, 实体关联性强, 在对序列的单个向量进行预处理时, 由于 Transformer Encoder 采用前馈传播特征向量的形式, 在单向传递的过程中所包含的语义信息逐层递减, 可能造成输出的特征向量语义不全的问题. MFE-BERT 模型在原有全连接的基础上, 将其中 11 个 Encoder 层中处理以后的信息输出至最上层, 此时模型拥有了输入序列不同抽象力度的特征信息, 通过 Concat 函数将 12 层中具有上下文语义信息的特征向量进行拼接, 最后输出一个具有多层级语义信息的特征向量. 该模型除了可以获取序列的上下文语义信息, 还可以将当前特征向量的语义信息更深层次地提取出来. 改进 BERT 模型结构图如图 3 所示.

特征向量具体运算过程如下.

向量$E_n = [e_1, e_2, \cdots, e_n]$进入 Encoder 层中经过线性变化得到查询矩阵 Q、表征上下文关系矩阵 K 以及内容矩阵 V, 通过缩放因子$\sqrt{d_k}$与 *Softmax* 函数之后, 对应的就是各个词之间的相互关联程度, 再点积内容矩

阵得到注意力分数值, 如式 (1):

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

MFE-BERT 模型中 Encoder 层使用了注意力机制, 因此式 (2) 为通过不同的线性变化获取 Q、K、V 矩阵, 点积对应的权重矩阵W_i^Q, W_i^K, W_i^V就能够得到每层注意力分值, 最后将其进行 Concat 拼接以后, 点乘附加的权重矩阵W_i^O, 即可获得具有上下文语义信息的特征向量 e_{ij}, 如式 (3):

$$head_i=Attention\left(QW_i^Q,KW_i^K,VW_i^V\right) \tag{2}$$

$$e_{ij}=Concat(head_1,\cdots,head_{12})W^o \tag{3}$$

式 (4) 是对 12 层 Encoder 中输出的特征向量进行拼接工作.

$$ce_i=Concat(e_{i1},e_{i2},\cdots,e_{i12}) \tag{4}$$

式 (5) 是对向量进行了全链接映射降维处理, 使得 MFE-BERT 模型训练出的特征向量维度与下游任务维度对应, 最终输出一个具有深层语义信息的特征向量.

$$x_i=\tanh(ce_i+b_i) \tag{5}$$

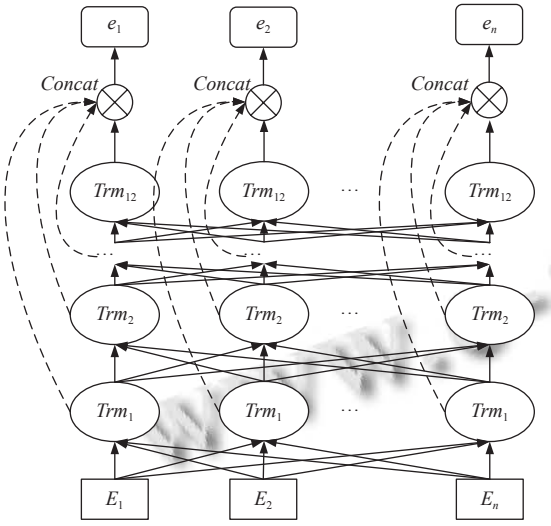


图 3 MFE-BERT 结构图

2.3.2 FNNAttention 机制

到目前为止, 对于心理医学长文本的实体识别与关系抽取仍然存在字符标签在长文本中不同一的问题. 研究证明长短期记忆网络能够有效联系的上下文文字个数余额为 200 个, 而我们心理医学研究文本段落远多于这个数字. 在长段落文本中, 位置相距较远的相同的心理医学实体概率性被算法赋予不同的实体标签, 存在长序列语义稀释问题, 这就使得模型正确率难以提高.

FNNAttention 的引进可以很好地解决这个问题, 其本质就是词向量的权重分配, 通过前向神经网络计算长距离词向量之间的语义关联程度, 自动学习调整语义权重, 强化词级关系. 我们通过 FNNAttention 可以有效利用长文本段落上下文之间的语义信息, 获取全局信息向量, 整合于当前特征向量进行计算, 把注意力主要分配给关键词, 在解决长文本中相同心理医学实体在不同语句中标注不一致问题的同时强化文本的词级关系. FNNAttention 模型图如图 4 所示.

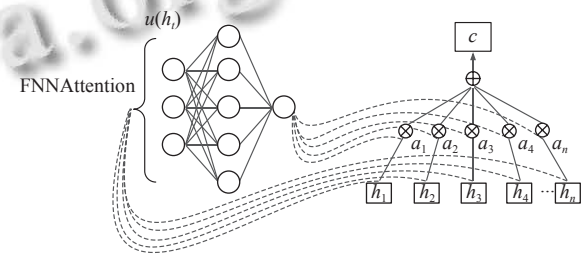


图 4 FNNAttention 结构图

具体运算细则如下.

式 (6) 是通过计算状态序列h_t的自适应加权平均来获取具有全局特征信息的向量c.

$$c=\sum_{i=1}^{T}\frac{\exp(u(h_t))}{\sum_{i=1}^{T}\exp(u(h_k))}h_t \tag{6}$$

其中, $\mu(\cdot)$是前向神经网络自学习函数, 仅通过状态序列h_t来学习, h_t是通过 BiLSTM 层输出的具有双向语义信息的状态序列表示, k为经过神经网络自学习得到的数值.

式 (7) 是对向量c与状态序列h_t进行的拼接操作, 通过非线性激活函数计算可以得到状态与向量的联合表示p_t.

$$p_t=\tanh[h_t,c] \tag{7}$$

式 (8) 和式 (9) 是对特征向量的降维处理, p_t通过一个全链接层得到低维向量g_t, g_t重复上一计算得到更低维向量z_t. z_t再次映射得到各字增强语义向量.

$$g_t=\tanh\left(W_pp_t+b_p\right) \tag{8}$$

$$z_t=\tanh(W_zp_t+b_z) \tag{9}$$

2.3.3 CRF 标签预测损失函数优化

CRF 模型能够在已给的输入序列基础上, 计算输

出序列的条件分布概率分布, 为了能够避免实体识别模型最后结果出现过拟合现象, 对 CRF 损失函数融入惩罚机制进行优化, 如式 (10):

$$L = -\left(S_t - \log\left(e^{S_1} + e^{S_2} + \cdots + e^{S_N}\right)\right) + \alpha \|\theta\|_2^2 \quad (10)$$

其中, S_t 是序列中真实标签路径的得分, $P_{\text{total}} = \left(e^{S_1} + e^{S_2} + \cdots + e^{S_N}\right)$为所有预测标签路径的总分, 在训练 CRF 模型时, 通常我们的目标是最小化损失函数, 因此加上一个负号; θ为实体识别模型中所有可训练参数, α为自定义交叉验证方法确定的超参数, 以此来惩罚模型中参数可以减少产生过拟合现象.

2.3.4 实体对齐

心理医学文本经过实体识别以及关系抽取后, 所形成的知识存在大量错误与冗余数据, 因此需要对抽取的知识进行数据清洗整合, 其主要工作即为实体对齐, 具体操作如下.

首先基于规则的方法将心理医学文本中标注的实体与同义名称进行信息映射, 组成同义实体库. 然后通过计算心理医学实体之间近似程度进行实体对齐, 包括实体文本相似度与实体语义相似度两部分.

(1) 实体文本相似度计算

文本相似度通过计算文本中两个实体之间的相同字符占比来判断是否为同一表述, 主要采用 Jaccard 系数来计算其相似程度, 如式 (11):

$$sim_t(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A + B + (A \cup B)|} \quad (11)$$

其中, $sim_t(A,B)$为实体 A 与实体 B 之间的文本相似度, $|A \cap B|$为两实体间相同字符的数量, $|A \cup B|$为两实体所有字符的数量和.

(2) 实体语义相似度计算

语义相似度利用实体在上下文语义中近似程度来计算两实体之间的相似度, 其方法是对实体间的词向量进行余弦计算. 余弦值越趋近于 1, 则向量角距离越小, 两实体间的近似程度越大, 如式 (12):

$$sim_s(A,B) = \cos(\theta) = \frac{\sum_i^n A_i \times B_i}{\sqrt{\sum_i^n (A_i)^2} \times \sqrt{\sum_i^n (B_i)^2}} \quad (12)$$

其中, $sim_s(A,B)$表示实体 A 与实体 B 之间的语义相似度, $A = [A_1, A_2, \cdots, A_n]$表示实体 A 的词向量集合, $B = [B_1, B_2, \cdots, B_n]$表示实体 B 的词向量集合.

由于部分不同实体文本过于相似, 单纯使用文本相似度很容易造成错误实体对齐, 如“躁郁症”与“抑郁症”两种病症文本相似度很高, 但为两种不同的实体. 因此, 本文结合两种相似度计算方法, 对文本相似度与语义相似度分别分配 0.4 与 0.6 的权重进行实体相似度计算, 如式 (13):

$$sim(A,B) = sim_t(A,B) \times 0.4 + sim_s(A,B) \times 0.6 \quad (13)$$

3 实验与分析

3.1 数据集与评价指标

(1) 心理医学数据集

心理医学领域缺乏标准型标注的数据资源, 数据集从友心理、有来医生、医学百科等 26 个心理医学网站采用正则表达式与 XPath 批量爬取相关文本数据, 对心理医学文本进行预处理清洗, 通过手工标注方式最终得到含有 3 927 个实体, 2 662 条关系的心理医学文本数据集. 其类别以及数量如下: 别称数量为 77, 不适部位数量为 166, 症状数量为 924, 检查数量为 313, 科室数量为 385, 并发症数量是 627, 易感人群数量为 170.

(2) ChineseBLUE 生物医学文本公开数据集

为了验证提出模型在公共数据集上对心理医学实体识别以及关系抽取的效果, 采用部分阿里公开的 ChineseBLUE 生物医学文本公开数据集进行测试. 其预定义类别同自建的心理医学数据集一致, 语料中训练集数据为 2 755 条文本, 验证集数据为 936 条文本, 测试集数据为 927 条文本.

普遍采用精确率 (*precision*)、召回率 (*recall*) 和 *F*1 值作为心理医学实体识别和关系抽取性能的评价指标, 本文主要以 *F*1 值作为性能指标判断, 指标计算公式如下所示:

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (14)$$

$$recall = \frac{TP}{TP + FN} \times 100\% \quad (15)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (16)$$

3.2 实验环境

心理医学知识图谱构建研究所有实验的具体环境如表 1 所示.

表 1 实验环境

实验环境	硬件配置
操作系统	Windows 10
CPU	Intel Core i7-10700 CPU
GPU	NVIDIA GeForce RTX 3070 (8 GB)
RAM	16 GB
TensorFlow	1.15
Python	3.7

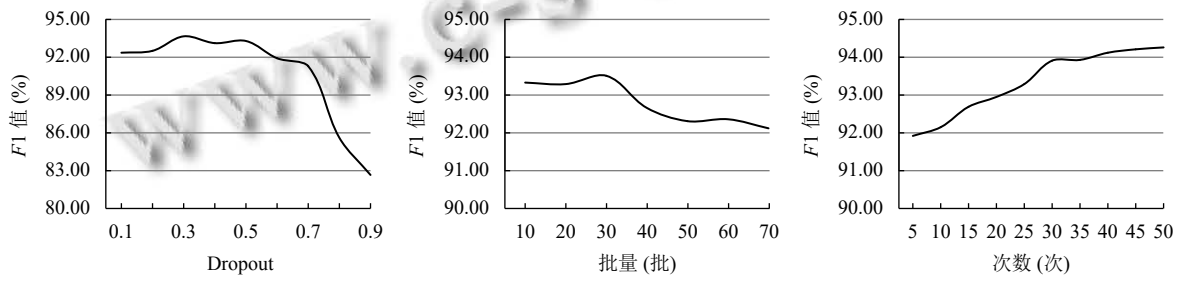
3.3 心理医学实体识别实验

3.3.1 实体标注

心理医学实体标注采用 BIO 的标志体系, 将搜集到的 3 927 个实体逐一标注, 其中 B (Begin) 代表一个心理医学实体的开始, I (Intermediate) 代表一个心理医学实体的中间部分, O (other) 表示其他部分. 实验标注 7 种实体类型, 包括心理疾病 (disease)、心理易患病人群 (susceptible_crowd)、心理疾病症状 (symptom)、心理疾病别名 (alternate_name)、不适部位 (pathogenic_site)、检查名称 (check)、科室名称 (department).

3.3.2 实体识别模型参数设置

对于心理医学文本数据集中输入序列较长的特点, 设置句子的最大长度为 300, 通过对模型在心理医学数据集进行丢弃率、批量大小、迭代次数等超参数的实验测试, 确定模型的参数. 实体识别的具体参数为: 选择 Adam 作为优化器, 学习率为 1E–4, 为防止模型训练过程中出现的过拟合现象, 引入丢弃率 (dropout) 设置为 0.3 (如图 5(a)), 批量大小设置为 32 (如图 5(b)), 迭代次数设置为 30 (如图 5(c)).



(a) 丢弃率 (b) 批量大小 (c) 迭代次数

图 5 不同超参数对模型性能的影响

在图 5(a) 中, 对 dropout 进行设置以防止模型出现过拟合现象, 在 dropout=0.3 时达到模型最优性能, 当 dropout>0.6 时, 模型性能过低, 这是由于丢弃率过高导致模型网络收敛速度减慢并且稳定性降低, 无法获取足量语义. 在图 5(b) 中, 批量大小在 30 附近模型性能达到最优, 降低模型批量大小之后难以达到收敛, 因此模型性能降低. 在图 5(c) 中, 模型 *F*1 值随着迭代次数的增多而增加, 当迭代次数大于 30 时, 模型 *F*1 值增加缓慢, 效率降低, 因此设置模型迭代次数为 30.

3.3.3 对比实验

为了验证模型在心理医学实体识别任务中的可行性, 将本模型在专业心理医学数据集和 ChineseBLUE 生物医学文本公开数据集上分别进行对比实验, 如表 2 所示.

对表 2 模型实验数据进行分析, 本文提出的 MFE-BERT-BiLSTM-FNNAttention-CRF 模型在自建的心理医学数据集测试中识别效果相比与其他方法更好, 精确率为 96.69%, 召回率为 92.85%, *F*1 值为 93.91%; 在 ChineseBLUE 生物医学公开数据集的实验中也取得了较好的成绩, 精确率为 91.16%, 召回率为 89.27%, *F*1 值为 89.52%, 说明其在心理医学实体识别任务中具有良好的适应性.

在对比实验中, 实验 1–6 在其他任务层相同的情况下, 分别对 BiLSTM 层与 CNN 层进行更换实验, 两数据集实验结果表明在实体识别任务中, BiLSTM 层可以充分利用上下文信息, 因此特征向量所具有的语义更全面, 对特征向量的处理效果更好; 将实验 7 与实验 8、实验 9 与实验 11 进行对比, 可以看出当选择 BERT 作为向量预训练模型使, 效果要优于 GPT-2, 这是由于 BERT 使用的双向的语言模型; 而 GPT-2 用的是单向语言模型, 因此 BERT 还可以在一定程度上获取上下文语义, 对于序列特征提取更充足; 对比实验 10 与实验 12, 将 MFE-BERT 作为预训练模型, 分别使用不同的注意力机制时, 在两数据集中取得了不同的效果, 在心理医学数据集中本文提出的改进模型达到最优, 其 *F*1 值为 93.91%, 在公开数据集中 *F*1 值为 89.52%, 仅次

于使用传统注意力机制的模型, 其原因是心理医学数据集是长文本数据, FNNAttention 在处理长文本任务中能够更好地强化语义信息, 而在公开数据集的短文本数据中, 传统 Attention 的权重分配计算更为合理, 对比两模型训练时间, 可以看出 FNNAttention 计算效率更优, 这是因为 FNNAttention 机制通过引入一个前馈神经网络来计算相似度, 减少了计算量. 从总体来看, 各模型在不同数据集之间的表现性能趋于一致, 且各模型在文本规范、格式固定, 经过数据清洗的心理医学数据集识别效果更好.

表 2 不同数据集实体识别对比结果

序号	模型	心理医学数据集				ChineseBLUE生物医学公开数据集			
		precision (%)	*recall* (%)	*F*1 (%)	*T* (s)	*precision* (%)	*recall* (%)	*F*1 (%)	*T* (s)
1	CNN-CRF[9]	87.19	86.95	87.06	29.1	81.56	81.33	81.45	53.3
2	BiLSTM-CRF[10]	88.29	86.65	87.46	33.6	82.93	81.96	82.16	60.2
3	BERT-CNN-CRF	93.82	89.46	91.67	93.7	87.23	87.57	87.39	136.9.
4	BERT-BiLSTM-CRF[11]	95.29	91.31	92.15	102.8	89.69	87.39	87.65	143.6
5	GPT2-CNN-CRF	91.61	89.39	90.46	267.3	86.92	86.76	86.83	293.9
6	GPT2-BiLSTM-CRF	92.13	90.25	91.16	281.2	87.65	87.89	87.76	322.4
7	BERT-BiLSTM-Attention-CRF[12]	95.96	91.36	92.95	136.5	90.35	88.46	88.76	175.2
8	GPT2-BiLSTM-Attention-CRF	91.83	90.21	91.01	302.9	87.29	86.65	86.96	357
9	GPT2-BiLSTM-FNNAttention-CRF	92.06	91.65	91.85	295.1	87.56	86.59	87.07	339.6
10	MFE-BERT-BiLSTM-Attention-CRF	96.33	92.12	93.56	145.3	90.51	89.03	89.76	190.5
11	BERT-BiLSTM-FNNAttention-CRF	96.16	92.15	93.32	132.6	90.19	88.63	89.21	167.3
12	**MFE-BERT-BiLSTM-FNNAttention-CRF**	**96.69**	**92.85**	**93.91**	**139.5**	**91.16**	**89.27**	**89.52**	**179.2**

综上, 对比实验中应用 MFE-BERT 模型的实体识别模型在两数据集中均能达到最优表现, 应用 FNN-Attention 模型的实体识别模型在心理医学长文本数据集中相对 Attention 模型表现更出色. MFE-BERT-BiLSTM-FNNAttention-CRF 模型对心理医学实体识别任务中可以很好提取序列的隐含特征, 在两种数据集中均表现出不错的性能, 能够有效提取心理医学相应实体.

3.3.4 实体识别实验结果

MFE-BERT-BiLSTM-FNNAttention-CRF 模型对心理医学数据集种的 7 种共 3 927 个标注实体进行性能测试实验, 实验结果如图 6 所示.

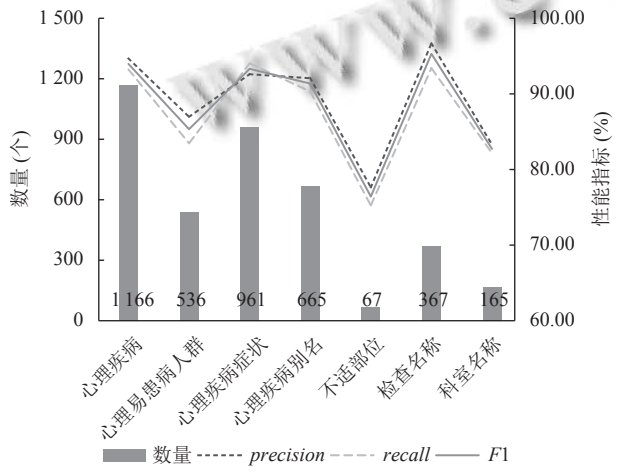


图 6 心理医学数据集实体识别数量及性能指标结果

从图 6 中可以看出, 心理疾病、心理疾病症状、心理疾病别名、检查名称这几种实体的 *F*1 值和精确率都在 90% 以上, 其中对检查名称实体识别的效果最好, 精确率为 96.63%, 召回率为 93.36%, *F*1 值为 95.25%, 这是因为在数据集中对其表述固定, 模型容易学习到其规则的描述方式; 心理易患病人群的实验 *F*1 值为 85.29%, 性能相对较差, 这是由于部分患病人群实体的表述种含有心理疾病症状, 比如“常见于焦躁男性”的表述, 将“焦躁”识别为症状, 这也是心理疾病症状实体召回率较高的原因; 对不适部位与科室名称实体识别的效果较差, 主要原因是在心理医学数据集种关于这两种实体的样本数较少, 达不到训练效果.

3.3.5 MFE-BERT 特征抽取层数量影响

为了研究 MFE-BERT 预训练模型提取不同数量 Encoder 层特征对实体识别结果的影响, 设置 6 组不同的提取层 L=[2, 4, 6, 8, 10, 12], 分别拼接第 6 层与第 12 层; 第 1 层、第 5 层、第 9 层、第 12 层, 以此类推, 每层间隔数量尽可能相近, 以保持特征分布的均匀性. 利用 MFE-BERT-BiLSTM-FNNAttention-CRF 模型在其他参数不变的基础上于自建的心理医学数据集进行实验, 实验结果如表 3 所示.

从表 3 中可以看出, 随着特征提取层数量的增加, 模型的性能逐步提升, 在提取 12 层全部特征时, 即本

模型所采用方案, 模型性能达到最佳, *F*1 值为 93.82%; 模型性能提升速率随着层数增加而变缓慢, 其原因是在拼接 Encoder 输出层的过程中, 拼接层次越多, 学习到所包含的不同语义信息就越少, 因此模型的性能提升就越小. 综上, MFE-BERT 模型所拼接的特征层越多, 就可以从心理医学文本数据集中获取越多的特征表示, 模型也将拥有更为丰富的表达能力, 从而提高模型的识别效果.

表 3 心理医学数据集不同数量特征抽取层实体识别结果 (%)

模型	选取层	*P*	*R*	*F*1
MFE-BERT-2	6, 12	96.32	92.37	93.52
MFE-BERT-4	1, 5, 9, 12	96.45	92.56	93.68
MFE-BERT-6	1, 3, 5, 7, 9, 12	96.53	92.69	93.79
MFE-BERT-8	1, 3, 5, 6, 7, 8, 10, 12	96.60	92.76	93.85
MFE-BERT-10	2, 3, 4, 5, 6, 8, 9, 10, 11, 12	96.66	92.80	93.89
MFE-BERT-12	1–12	96.69	92.85	93.91

3.4 心理医学关系抽取实验

3.4.1 关系类别

将心理医学关系进行分类, 主要包括 7 大类, 在自建心理医学数据集中, 其类别以及数量如下: 别称 (is_alternate_name) 数量为 77, 不适部位 (is_pathogenic_site) 数量为 166, 症状 (has_symptom) 数量为 924, 检查 (is_check) 数量为 313, 科室 (is_department) 数量为 385, 并发症 (accompany_with) 数量是 627, 易感人群 (is_susceptible_crowd) 数量为 170, 共 2 662 条心理医学实体关系.

3.4.2 参数设置

根据实验调试, 对 MFE-BERT-CNN-FNNAttention 关系抽取模型中所涉及的重要参数设置如表 4 所示.

表 4 关系抽取模型参数设计

参数	参数值
max_len (最大长度)	300
batch_size (批次大小)	32
优化器	Adam
激活函数	ReLU
epoch (迭代次数)	256
learning rate (学习率)	1E–4
dropout (丢弃率)	0.5

3.4.3 对比实验

为了验证模型在心理医学关系抽取任务种的可行性, 将本模型在自建心理医学数据集和 ChineseBLUE 生物医学文本公开数据集上分别进行对比实验, 如表 5 所示.

表 5 不同数据集关系抽取对比结果

序号	模型	心理医学数据集				ChineseBLUE生物医学公开数据集			
		P (%)	*R* (%)	*F*1 (%)	*T* (s)	*P* (%)	*R* (%)	*F*1 (%)	*T* (s)
1	Word2Vec-CNN[16]	83.16	82.02	82.58	67.3	81.65	79.95	80.79	106.2
2	Word2Vec-BiLSTM	82.92	81.31	82.10	85.6	81.23	79.12	79.92	136.4
3	BERT-CNN[16]	87.90	85.69	86.78	262.8	86.06	84.69	85.36	372.2
4	BERT-BiLSTM	87.25	85.12	86.17	293.7	85.36	84.23	84.79	416.9
5	Word2Vec-CNN-Attention[19]	83.85	82.33	83.08	132.1	82.52	81.27	81.89	179.5
6	Word2Vec-BiLSTM-Attention[17]	83.61	81.64	82.62	147	81.61	80.83	81.23	196.3
7	BERT-CNN-FNNAttention	88.23	86.91	87.36	331.6	85.63	86.12	85.89	429.7
8	BERT-CNN-Attention	87.67	86.46	87.06	356.2	86.75	86.26	86.51	459
9	MFE-BERT-CNN-Attention	88.65	87.56	88.19	396.9	87.85	87.01	87.43	506.1
10	**MFE-BERT-CNN-FNNAttention**	**89.29**	**87.69**	**88.32**	**362.5**	**87.49**	**86.67**	**87.12**	**479.6**

对关系抽取实验结果进行分析, 本文所提出的 MFE-BERT-CNN-FNNAttention 心理医学关系抽取模型在自建的心理医学数据集取得了最好的效果, 精确率达到 89.39%, 召回率为 88.52%, *F*1 值为 88.95%; 在生物医学公开数据集中效果显著, 精确率为 87.49%, 召回率为 86.67%, *F*1 值为 87.12%, 可以看出所改进的模型性能相对于基础模型都有了一定的提升, 能够有效对心理医学文本数据进行抽象建模.

在两个数据集的关系抽取实验中, 将实验 1–4 进行对比, BERT 预训练模型的效果要优于 Word2Vec 预训练模型, 且 CNN 模型对关系特征的抽取能力更胜于 BiLSTM 模型; 将实验 1 与实验 5, 实验 2 与实验 6 进行比较, 可以看出 Attention 机制对于关系抽取实验的精确度起着积极作用, 主要因为注意力机制进行了权重分配, 将更多注意力给心理医学关系; 对比实验 8 与实验 9, 预训练模型从 BERT 更换为本文提出的改进的 MFE-BERT, 在心理医学数据集中 *F*1 值分别为 87.06%、88.19%, 说明改进的 BERT 利用层级拼接方

式学习到了序列的深度词级关系信息; 再次将实验 7 与实验 8, 实验 9 与实验 10 进行对比, 由于受文本长度的影响, FNNAttention 机制仅在长文本数据集中能发挥其优势, 在心理医学数据集中 FNNAttention 表现比传统 Attention 更好, 在用时相对少的前提下 $F1$ 值有了一定程度的提升, 在公开数据集中 FNNAttention 性能的提升仅次于 Attention, 但是仍然比不加注意力机制的模型效果好.

综上, MFE-BERT-CNN-FNNAttention 模型在心理医学关系抽取任务中表现良好, 所改进的 MFE-BERT 预处理模型能够很好地适应不同数据集的心理医学关系抽取任务.

3.4.4 关系抽取实验结果

MFE-BERT-CNN-FNNAttention 模型对心理医学数据集进行关系抽取, 对 7 种关系类别进行性能指标测试, 实验结果如图 7 所示. 通过实验可以看出, 心理疾病并发症和心理易患病人群两种关系的性能指标相对较低, $F1$ 值分别为 83.61% 和 83.95%, 这是由于心理疾病并发症关系描述表达容易与疾病症状进行混淆, 且在数据集中对心理易患病人群关系的语料覆盖不足, 导致这两种关系抽取识别率不高. 疾病所属科室关系抽取效果最好, 其 $F1$ 值达到了 91.29%, 这是因为在实体识别任务中对应中文词组较少, 且标注正比例样本较多, 因此疾病所属科室关系可以被极大部分提取出来.

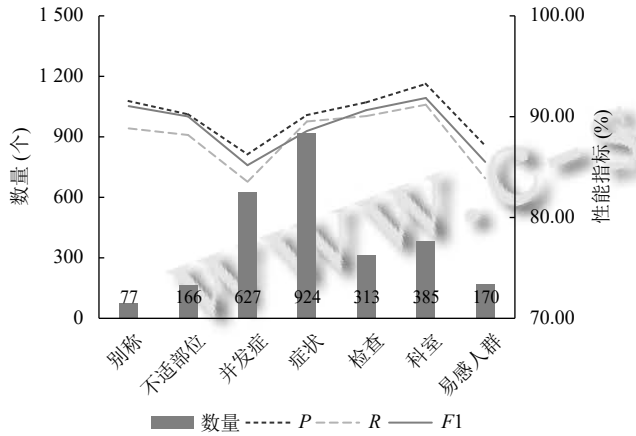


图 7 心理医学数据集关系抽取数量及性能指标结果

3.4.5 MFE-BERT 特征抽取层数量影响

为了研究 MFE-BERT 预训练模型提取不同数量 Encoder 层特征对关系抽取结果的影响, 设置 6 组不同的提取层 L=[2, 4, 6, 8, 10, 12], 其拼接方式同实体识别特征抽取实验一致. 利用 MFE-BERT-CNN-FNNAttention 模型在其他参数不变的基础上于自建的心理医学数据集进行实验, 实验结果如表 6 所示.

表 6 心理医学数据集不同数量特征抽取层关系抽取结果 (%)

预训练模型	选取层	P	R	$F1$
MFE-BERT-2	6, 12	88.62	87.09	87.81
MFE-BERT-4	1, 5, 9, 12	88.76	87.32	87.93
MFE-BERT-6	1, 3, 5, 7, 9, 12	88.83	87.58	88.15
MFE-BERT-8	1, 3, 5, 6, 7, 8, 10, 12	88.99	87.56	88.23
MFE-BERT-10	2, 3, 4, 5, 6, 8, 9, 10, 11, 12	89.06	87.66	88.26
MFE-BERT-12	1–12	89.29	87.69	88.32

从表 6 中可以看出, 随着特征提取层数量的增加, 关系抽取模型的性能逐步提升, 同实体识别模型表现一致, 在提取所有层 Encoder 特征时, 模型性能达到最佳, $F1$ 值为 88.32%. 实验说明 MFE-BERT 模型所拼接的特征层越多, 关系抽取模型对于整个文本的理解程度就越深入, 所学习到实体间的语义特征就越多, 从而提高模型的关系抽取效果.

3.5 心理医学知识图谱可视化

心理医学知识图谱可视化主要通过数据层提取信息以及前端视图层展示关系两部分组成. 数据层通过 MFE-BERT-BiLSTM-FNNAttention-CRF 模型进行实体识别, MFE-BERT-CNN-FNNAttention 模型进行关系抽取, 两复合模型都是基于多层特征提取能力对 BERT 进行改进处理, 同时应用前向神经网络注意力机制. 知识图谱实体内容简洁但关系数据量巨大, 若使用关系型数据库如 MySQL 会产生大量的表结构, 造成数据冗余浪费大量存储空间, 因此对心理医学实体和关系存储时采用图数据库, 在提高查询效率的同时保证合理的存储空间, 前端视图层通过 Neo4j 图数据库浏览器可视化进行展示. 部分心理医学知识图谱可视化效果如图 8 所示.

4 结束语

本文提出的心理医学实体识别模型以及心理医学实体关系抽取模型, 将心理医学数据集在两模型上进行训练提取心理医学实体及关系数据, 构建成一个领域内容全面且可视化的心理医学知识图谱. 通过应用 FNNAttention 机制以及对 BERT 基础预训练模型的改进, 以加强模型的语义特征提取能力, 同时改善在长段落文本中语义稀释的问题, 在心理医学数据集、Chinese-BLUE 公开数据集实验中均取得了较好的效果, 性能指标 $F1$ 值均在 85% 以上.

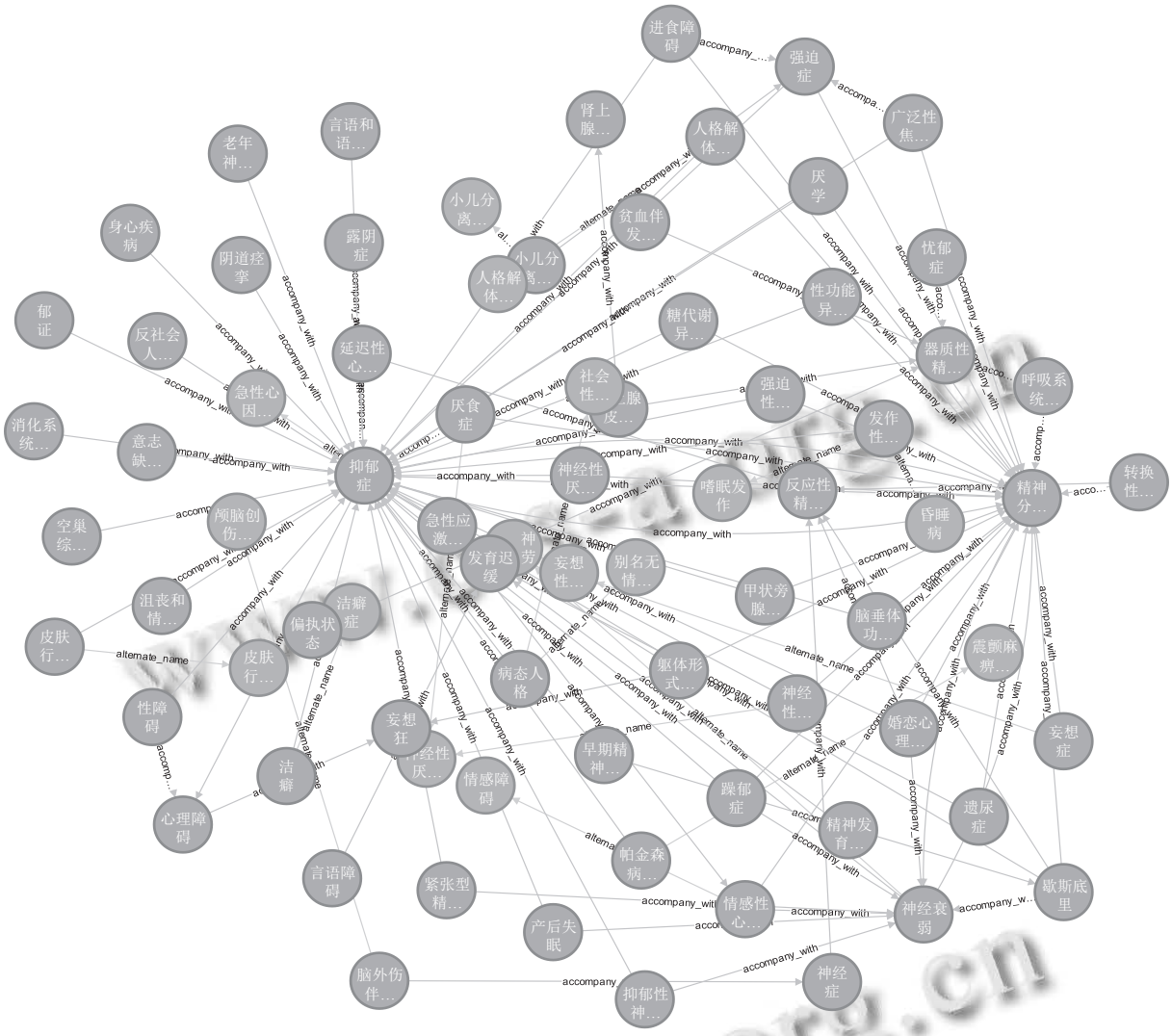


图 8 心理医学知识图谱

后续研究还需从以下两个方面进一步优化.

①以满足心理医学特定应用背景的实际需求为前提, 考虑模型性能与训练效率的平衡问题, 设计高效的实体识别及关系抽取模型.

②尝试拓展实验样本的丰富性与多样性, 探索多任务学习方法, 针对不同领域进行相应知识图谱构建任务, 提高模型普适性.

参考文献

1 Wang YZ, Jia YT, Liu DW, *et al*. Open Web knowledge aided information search and data mining. Journal of Computer Research and Development, 2015, 52(2): 456–474.

2 Wang Q, Mao ZD, Wang B, *et al*. Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724–2743. [doi: 10.1109/TKDE.2017.2754499]

3 徐春, 李胜楠. 融合 BERT-WWM 和指针网络的旅游知识图谱构建研究. 计算机工程与应用, 2022, 58(12): 280–288.

4 Martinez-Rodriguez JL, Lopez-Arevalo I, Rios-Alvarado AB. OpenIE-based approach for knowledge graph construction from text. Expert Systems with Applications, 2018, 113: 339–355. [doi: 10.1016/j.eswa.2018.07.017]

5 黄梦醒, 李梦龙, 韩惠蕊. 基于电子病历的实体识别和知识图谱构建的研究. 计算机应用研究, 2019, 36(12): 3735–3739. [doi: 10.19734/j.issn.1001-3695.2018.07.0414]

6 廖开际, 黄琼影, 席运江. 在线医疗社区问答文本的知识图谱构建研究. 情报科学, 2021, 39(3): 51–59, 75. [doi: 10.13833/j.issn.1007-7634.2021.03.008]

7 Raffel C, Ellis DPW. Feed-forward networks with attention can solve some long-term memory problems. arXiv:

1512.08756, 2015.
8 Li J, Sun AX, Han JL, *et al*. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50–70. [doi: 10.1109/TKDE.2020.2981314]
9 Collobert R, Weston J, Bottou L, *et al*. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011, 12: 2493–2537.
10 Habibi M, Weber L, Neves M, *et al*. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics, 2017, 33(14): i37–i48. [doi: 10.1093/bioinformatics/btx228]
11 郭知鑫, 邓小龙. 基于 BERT-BiLSTM-CRF 的法律案件实体智能识别方法. 北京邮电大学学报, 2021, 44(4): 129–134. [doi: 10.13190/j.jbupt.2020-241]
12 任媛, 于红, 杨鹤, 等. 融合注意力机制与 BERT+BiLSTM+CRF 模型的渔业标准定量指标识别. 农业工程学报, 2021, 37(10): 135–141.
13 Lakel K, Bendella F, Benkhadda S. Named entity recognition for psychological domain: Challenges in document annotation for the Arabic language. Proceedings of the 1st International Conference on Embedded & Distributed Systems (EDiS). Oran: IEEE, 2017. 1–5.
14 李冬梅, 张扬, 李东远, 等. 实体关系抽取方法研究综述. 计算机研究与发展, 2020, 57(7): 1424–1448.
15 Jain A, Pennacchiotti M. Open entity extraction from Web search query logs. Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: ACM, 2010. 510–518.
16 陆晓蕾, 倪斌. 基于预训练语言模型的 BERT-CNN 多层级专利分类研究. 中文信息学报, 2021, 35(11): 70–79.
17 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 (Volume 2: Short Papers). Berlin: ACL, 2016. 207–212.
18 刘峰, 高赛, 于碧辉, 等. 基于 Multi-head Attention 和 Bi-LSTM 的实体关系分类. 计算机系统应用, 2019, 28(6): 118–124. [doi: 10.15888/j.cnki.csa.006944]
19 姚宁, 张淼, 陈宏涛. 基于 CNN-Attention 算法的精神分裂症分类. 电子设计工程, 2022, 30(10): 55–61. [doi: 10.14022/j.issn1674-6236.2022.10.012]
20 宋伟, 张游杰. 基于环境信息融合的知识图谱构建方法. 计算机系统应用, 2020, 29(6): 121–125. [doi: 10.15888/j.cnki.csa.007424]
21 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述. 计算机学报, 2017, 40(6): 1229–1251.

(校对责编: 牛欣悦)