计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.009640]

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

面向生物医学命名实体识别和规范化的多粒度特征融合①

刘　彤, 石昌岭, 倪维健

(山东科技大学 计算机科学与工程学院, 青岛 266590)
通信作者: 倪维健, E-mail: niweijian@sdust.edu.cn

摘　要: 为了从生物医学文献中提取丰富的实体信息及其规范化表达, 提出了一种面向生物医学命名实体和规范化的多粒度特征融合方法 (multi-granularity feature fusion approach for biomedical named entity recognition and normalization, MGFFA). 通过整合字符级、词级、概念级的文本信息, 显著增强了模型的学习能力. 同时还包含一个用于存储和综合不同层次信息的记忆库, 以实现对实体及其规范化标签间复杂关系的深入理解. 通过预训练模型的配合使用, MGFFA 不仅捕捉了文本的粗粒度语义表示, 还细致分析了构词层面的特征, 从而全面提升了对长跨度实体的识别准确率. 在 NCBI 和 NC5CDR 数据集上的实验结果显示, 该模型在总体上优于其他基线模型.

关键词: 生物医学命名实体识别; 生物医学命名实体规范化; 多任务学习; 记忆网络

引用格式: 刘彤,石昌岭,倪维健.面向生物医学命名实体识别和规范化的多粒度特征融合.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9640.html

Multi-granularity Feature Fusion for Biomedical Named Entity Recognition and Normalization

LIU Tong, SHI Chang-Ling, NI Wei-Jian

(School of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

Abstract: To extract rich entity information and normalized expressions from biomedical literature, this study proposes a multi-granularity feature fusion approach for biomedical named entity recognition and normalization (MGFFA). By integrating character-level, word-level, and concept-level textual information, the model significantly enhances its learning capability. It also incorporates a memory bank for storing and synthesizing information from different levels to achieve a deeper understanding of the complex relationships between entities and their normalized labels. With the integration of pre-trained models, MGFFA captures not only coarse-grained semantic representations of text but also conducts detailed analysis at the morphological level, thereby comprehensively improving the recognition accuracy of long-span entities. Experimental results on the NCBI and NC5CDR datasets demonstrate that the model outperforms other baseline models overall.

Key words: biomedical named entity recognition; biomedical named entity normalization; multi-task learning; memory network

生物医学命名实体识别 (biomedical named entity recognition, BioNER)[1]和命名实体规范化 (biomedical named entity normalization, BioNEN)[2]是生物医学文本挖掘的关键任务, 分别用于从文献中识别实体和将实体映射到标准化概念. 这些基础任务对医学研究和临床决策至关重要. 由于传统的流水线模型存在错误传

① 基金项目: 科技创新 2030—“新一代人工智能”重大项目 (2022ZD0119500); 山东省自然科学基金 (ZR2022MF319); 山东科技大学青年教师教学拔尖人才培养基金 (BJ20211110)
收稿时间: 2024-04-01; 修改时间: 2024-04-29; 采用时间: 2024-05-09; csa 在线出版时间: 2024-09-27

播和灵活性不足的问题, 因此联合模型越来越受到重视, 它通过一个统一框架同时处理两个任务, 能减少错误传播并提升处理效率. 然而, 实际应用中联合模型仍然面临任务关联性和组件独立性等挑战.

生物医学命名实体识别和规范化是两个互补的过程, 它们共同构成了从生物医学文本中提取和利用关键信息的基础. 为此, 本文设计了一个联合标注方案, 可以同时编码实体信息和相应的概念信息, 将两种信息结合起来共同标注生物医学文本中的实体. 在此基础上, 本文提出了一种面向生物医学命名实体识别和规范化地多粒度融合模型 (multi-granularity feature fusion approach for biomedical named entity recognition and normalization, MGFFA). MGFFA 模型通过整合字符级、词级、概念级的文本信息, 显著增强了模型的学习能力. 模型包含一个动态更新的记忆库, 用于存储和综合不同层次信息, 以实现对实体及其规范化标签间复杂关系的深入理解. 通过预训练模型 BioBERT 和 CharBERT 的配合使用, MGFFA 不仅捕捉了文本的粗粒度语义表示, 还细致分析了构词层面的特征, 并融合了概念级信息, 从而全面提升了对长跨度实体的识别准确率.

1 相关工作

1.1 生物医学命名实体识别

近年来, 生物医学命名实体识别任务主要围绕循环神经网络[3]、长短期记忆网络[4]、条件随机场[5]等模型结构展开. Guan 等人[6]提出了一个名为 BioByGANS 的生物医学命名实体识别模型, 该模型结合了前缀特征和注意力图判别融合技术, 旨在提高命名实体识别任务的性能. 通过实验验证, 该方法在多个数据集上均表现出了优越的性能, 为生物医学领域的命名实体识别任务提供了新的解决方案. Malmasi 等人[7]构建了一个 MultiCoNER 的大规模多语言数据集, 用于命名实体识别. MultiCoNER 数据集的构建旨在提供一个广泛覆盖多语言的资源, 以促进复杂命名实体识别模型的发展和评估. 该数据集为研究人员提供了丰富的语言材料, 以支持多语言命名实体识别任务的进展. Li 等人[8]针对 NER 任务提出了一种新的方法 W2NER, 将其建模为词与词之间的关系分类任务. 实验证明, 该方法在多个基准数据集上均表现出优于基线模型的性能, 为命名实体识别任务的研究提供了新的思路和解决方案.

1.2 生物医学命名实体规范化

生物医学命名实体规范化是自然语言处理领域的另一个关键任务. 传统方法依赖于领域特定的词典[9]、手工制定的规则以及特征工程[10,11], 而最近深度学习方法因其强大的特征自动提取能力被广泛应用. Mehmood 等人[12]提出了一种多任务模型 MTM-CNN, 用于提高生物医学实体规范化任务的性能, 强调了知识转移技术在生物医学 NER 中的重要性, 以及利用外部资源和领域知识的关键性. 此外, 该模型还针对每个特定数据集进行了微调, 以将其从通用特征表示定制为专用特征表示. Sung 等人[13]提出了一种用于实体规范化的模型 BioSyn, 通过纳入同义词信息来增强生物医学实体的表示. 通过边缘化同义词的影响, 模型可以生成更健壮和准确的生物医学实体表示. Peng 等人[14]也提出了一种用于实体规范化的模型 IA-BIOSYN. 该模型基于交互式同义词边缘化的方法, 将多个语义特征整合在一起, 以增强模型捕获实体同义性的能力. 该模型在 NCBI-Disease、BC5CDR-Disease 和 BC5CDR-Chemical 数据集上验证了其优秀的性能表现, 并优于其他基线模型.

1.3 联合生物医学命名实体识别和规范化

随着技术的发展, 越来越多的研究开始使用联合模型来克服流水线模型的弊端. Zhou 等人[15]研究提出了一种名为 MTAAL 的模型, 结合了多任务学习和对抗性主动学习. MTAAL 模型通过对抗性主动学习, 模型能够有效地利用标签有限的数据来训练, 并且在命名实体识别和规范化任务上取得了显著的性能提升. 在 Zhou 等人[16] 的进一步研究中, 提出了一个名为 E2EMERN 的端到端渐进式多任务学习模型. 该模型采用了渐进式的多任务学习方法, 通过逐步学习不同任务的相关知识, 逐步提升模型的性能. 实验结果表明, 该模型在医学命名实体识别和规范化任务上取得了显著的性能提升. Ji 等人[17] 提出的 NeubRN 模型采用了神经过渡式联合框架, 将生物医学命名实体识别和规范化任务转换为一系列动作预测任务, 并通过注意力机制充分利用词汇表中每个候选概念的详细信息, 有效地支持了命名实体是被任务. 另一方面, Zhao 等人[18]提出了一种神经多任务学习模型 MTL-MEN&MER_feedback, 用于医学命名实体识别和规范化的联合建模. 该方法采用循环神经网络 (RNN) 和门控循环单元 (GRU), 并利用多任务学习策略来整合实体识别和规范

化任务, 同时引入了共享的表示层和任务特定的输出层. 这种方法在多个医学领域的基准数据集上取得了显著的性能提升.

2 方法

2.1 联合标注方案

为了从文本中提取精确且一致的生物医学信息, 生物医学命名实体识别和规范化任务的融合显得尤为重要. 本文提出了一种独特的标注方案, 旨在同时编码实体提及及其对应的规范概念, 将 BioNER 和 BioNEN 的联合任务转化为一个序列标注问题[19]. 联合标注方案不仅能够消除级联错误, 还加强了 BioNER 和 BioNEN 两个任务的紧密结合, 使得模型能够更好地识别出实体, 并将其准确地映射到概念标签上.

图 1 给出了所提出的联合标注方案的示例. 这一方案中, B-Disease、E-Disease 和 S-Disease 等实体标签用于指示命名实体提及的位置和类型; 而 D013921 和 D06493 等概念标签则作为来自参考词典的概念标识符. 通过为实体的每个 Token 分配一个概念标签, 并将其与相应的实体标签进行连接, 这种方法成功地集成了实体位置信息和概念标识符于单一的联合标签中. 以图 1 中词元 HIT 为例, 联合标签 B-D013921 表示该词元是在提及概念 D013921 的开头. 同时, 为了使联合标签更加的简洁, 本节省去了类型 Disease, 因为其已经包含在了概念标识符中了.

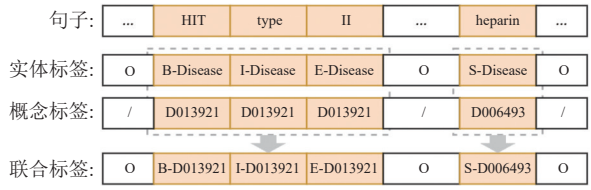

图 1 联合标注方案示例

2.2 多粒度特征融合方法 (MGFFA)

本文提出的 MGFFA 模型用于生物医学命名实体识别和规范化的联合抽取任务. 如图 2 所示, MGFFA 模型主要由以下几个模块组成: 单词表示模块、字符表示模块、多粒度融合模块、记忆模块和标签解码模块, 本节将对这几个模块进行详细介绍.

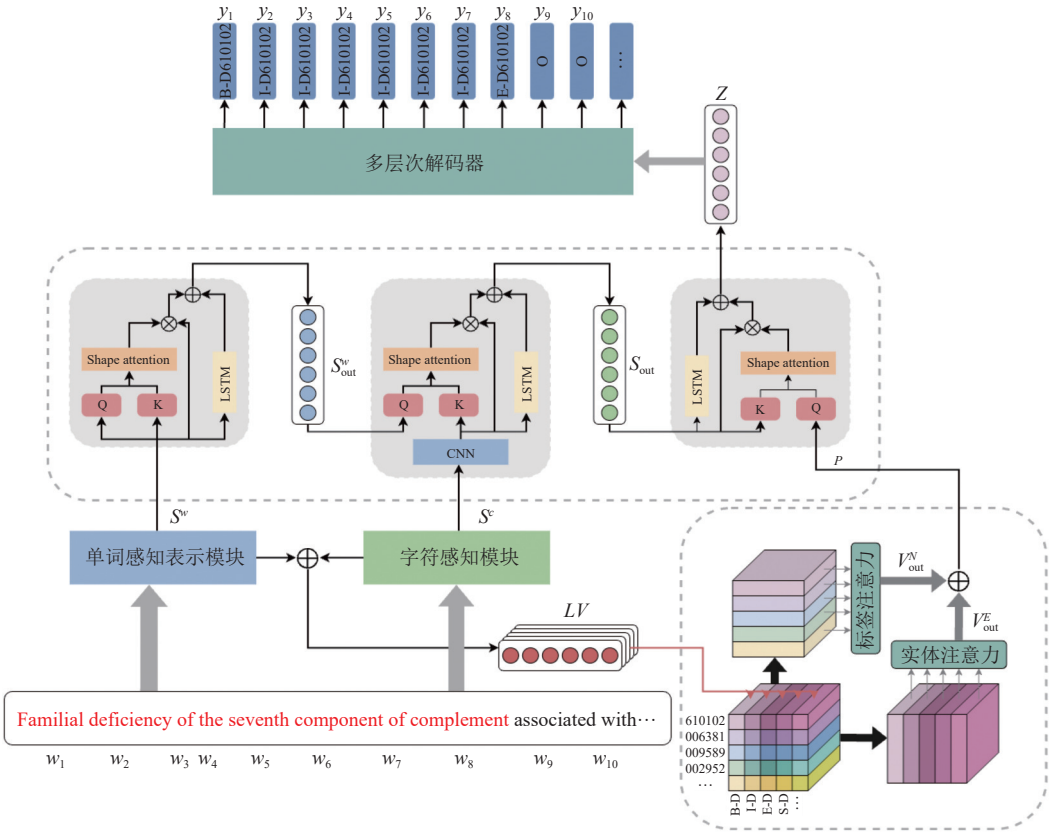

图 2 MGFFA 整体模型架构

2.2.1 单词和字符表示模块

成功的神经序列标注模型很大程度上依赖于输入序列的综合表示. 考虑到基于 BERT[20] 的预训练语言模型在多种任务上所展现的卓越性能[21,22], 本文选择了 BioBERT[23]作为输入句子的基础编码器. BioBERT 是一种专门在生物医学文本上预训练的 BERT 模型, 能够通过捕获文本中单词间的上下文依赖关系, 构建出丰富的单词表示. 然而, 生物医学文本中的单词构成通常遵循特定的规则, 如前缀、后缀和词根对于理解生物医学术语的语义至关重要. 传统的 BERT 系统, 包括 BioBERT 在内, 往往难以充分捕获这种构词学特征. 鉴于此, 本文进一步引入了一种字符感知的预训练语言模型 CharBERT[24], 它通过融合字符级信息, 产生更为细粒度的单词表示.

在 MGFFA 模型中, 每个句子被赋予两种类型的表示: 单词感知表示和字符感知表示. 具体来说, 对于一个包含n个单词的句子$S=(w_1,\cdots,w_n)$, 预训练得到的单词感知表示和字符感知表示分别为$S=(w_1,\cdots,w_n)$和$S^c=(x_1^c,\cdots,x_n^c)$. 这里, x_t^w和x_t^c $(t=1,\cdots,n)$分别表示第t个词的词和字符感知编码向量.

2.2.2 记忆模块

记忆模块是一种外部的记忆存储器, 可以被显式地读写[25,26], 使得模型能够灵活地存储和检索信息. 记忆模块使用更加轻量化的设计, 在联合标签矩阵生成以后, 将其在纵向和横向方面融合, 生成实体记忆部分和标签记忆部分.

(1) 联合标签矩阵. 联合标签是由实体标签和规范化标签组成, 如图 3 所示. 例如, 联合标签 B-D610102 表示编号 610102 的疾病实体开头, 可以将其拆分为实体表示 B-D 和规范化表示 610102, 并嵌入到坐标为(B-D,610102)的联合标签矩阵中.

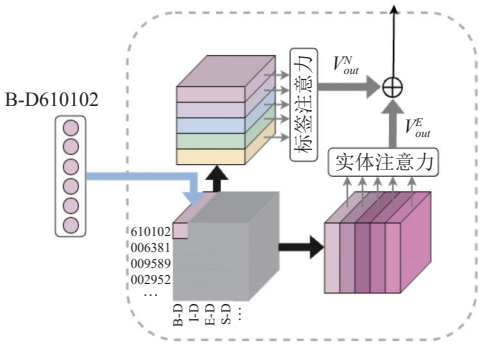

图 3 记忆标签

本质上联合标签矩阵M是一个存储向量的三维矩阵, 横轴表示的是实体标签, 纵轴表示的是规范化标签, 深度表示的是存储向量的维度. 设实体标签集$L^E=\{l_1^E,\cdots,l_H^E\}$, 其中$H$表示实体标签数量. 设规范化标签集$L^N=\{l_1^N,\cdots,l_K^N\}$, 其中$K$表示规范化标签的数量. 通过联合标签$y_i\,(i=1,\cdots,N)$可得到该单词在联合标签矩阵$M$中的坐标$(h,k)$, 然后将该标签的向量插入坐标$(h,k)$中. 对于每个标签的向量, 是由 BioBERT 和 CharBERT 生成的单词和字符感知编码连接得到:

$$LV=S^w\oplus S^c \tag{1}$$

其中, $\oplus$表示向量连接.

(2) 实体记忆部分. 联合标签矩阵在纵向上形成H个实体表示$V_i^E\in\mathbb{R}^{K\times d},i\in(1,\cdots,H)$. 在实体方向上, 当一组向量同属于一个实体标签, 但分别属于不同的规范化标签时, 进行自注意力查询的能够使模型聚焦于同一实体内的不同规范化标签. 对于每组实体分别进行自注意力计算, 可以表示为:

$$\begin{cases} T_i^E=Softmax\left(\dfrac{1}{\sqrt{d_k}}V_i^EW_i^Q(W_i^K)^{\mathrm{T}}(V_i^E)^{\mathrm{T}}\right)\\ V_{\mathrm{out}i}^E=\left(\alpha_iI_i+\beta_iT_i^E-\gamma_iC\right)V_i^E \end{cases} \tag{2}$$

其中, $W_i,\alpha_i,\beta_i,\gamma_i$都是可训练的参数, $\alpha_i,\beta_i,\gamma_i$初始化是均设为 1. C是常量.

(3) 标签记忆部分. 联合标签矩阵在横向上形成K个规范化表示$V_j^N\in\mathbb{R}^{H\times d},j\in(1,\cdots,K)$. 同样的, 在每个规范化标签的向量上进行自注意力操作, 模型可以更好地理解同一规范化标签下不同实体标签之间的关系. 对于每个规范化表示进行自注意力计算, 可以表示为:

$$\begin{cases} T_j^N=Softmax\left(\dfrac{1}{\sqrt{d_k}}V_j^NW_j^Q(W_j^K)^{\mathrm{T}}(V_j^N)^{\mathrm{T}}\right)\\ V_{\mathrm{out}j}^N=\left(\alpha_jI_j+\beta_jT_j^N-\gamma_jC\right)V_j^N \end{cases} \tag{3}$$

其中, $W_j,\alpha_j,\beta_j,\gamma_j$都是可训练的参数, $\alpha_j,\beta_j,\gamma_j$初始化时均设为 1. C是常量.

再对$V^E=\{V_1^E,\cdots,V_H^E\}$和$V^N=\{V_1^N,\cdots,V_K^N\}$进行 Attention 操作后分别得到$V_{\mathrm{out}}^E$和$V_{\mathrm{out}}^N$, 然后将这两个向量表示进行线性转换后相加, 得到标签概念粒度的向量表示P.

$$T^E=Softmax\left(\frac{1}{\sqrt{d_k}}V_{\mathrm{out}}^EW_1^Q(W_1^K)^{\mathrm{T}}(V_{\mathrm{out}}^E)^{\mathrm{T}}\right) \tag{4}$$

$$V_{\text{out}}^{E}=\left(\alpha_1 I_1+\beta_1 T^{E}-\gamma_i C\right)V_{\text{out}}^{E} \tag{5}$$

$$T^{N}=Softmax\left(\frac{1}{\sqrt{d_k}}V_{\text{out}}^{N}W_2^{Q}(W_2^{K})^{\text{T}}(V_{\text{out}}^{N})^{\text{T}}\right) \tag{6}$$

$$V_{\text{out}}^{N}=\left(\alpha_2 I_2+\beta_2 T^{N}-\gamma_j C\right)V_{\text{out}}^{N} \tag{7}$$

$$P=V_{\text{out}}^{E}\oplus V_{\text{out}}^{N} \tag{8}$$

2.2.3 多粒度特征融合模块

标签概念向量、词向量和字符向量分别代表生物医学文本中的标签、单词和字符, 属于 3 种不同粒度的信息. 这些向量能捕捉数据的不同尺度和层次信息, 对生物医学命名实体识别和规范化任务具有不同的重要性. 本文借鉴简化版 Transformer[27]的概念, 提出了一种并行 Transformer 模块, 优化利用多粒度信息. 此并行块通过 Transformer 优化和参数调整, 有效提升模型运行效率, 同时保持高性能, 证明了在增加效率的同时, 模型性能未受影响.

(1) 对于并行 Transformer 模块. 假定输入序列 $X_{\text{in}}\in\mathbb{R}^{T\times d}$ 包含 T 个 Token, emb 维度为 d, H 个头的多头 Attention 矩阵 $\widetilde{MHA}$ 任一 Attention 矩阵可表示为:

$$A_h(X)=Softmax\left(\frac{1}{\sqrt{d_k}}XW_h^{Q}(W_h^{K})^{\text{T}}X^{\text{T}}\right) \tag{9}$$

其中, $h\in(1,\cdots,H)$, $W_h^{Q},W_h^{K}\in\mathbb{R}^{d\times d/H}$.

构建一个新的 Shaped Attention 矩阵, 使之等价于单位矩阵. 这样带来的效果就是注意力权重分配机制中, Token 更多的是关乎自身, 非常类似于残差网络的机制, 因此信号可以在很深的网络中良好传播, 更容易训练. 可以表示为:

$$\widetilde{Attn}_h(X)=(\alpha_h I_T+\beta_h A_h(X)-\gamma_h C)X_h \tag{10}$$

其中, $X_h\in\mathbb{R}^{T\times d/H}$, $\alpha_h,\beta_h,\gamma_h$ 都是可训练的参数, 初始化时均设为 1. C 是常量.

对于 $\widetilde{MHA}$:

$$\widetilde{MHA}(X)=concat\left(\widetilde{Attn}_1(X),\cdots,\widetilde{Attn}_H(X)\right) \tag{11}$$

其中, X_h 是 $X\in\mathbb{R}^{T\times d}$ 的一部分, 即 $X=concat(X_1,\cdots,X_H)$.

在生物医学文本中常使用大量缩写词, 很多时候缩写词的完整含义只能通过上下文来确定. 为了更好地捕捉生物医学文本序列中的长时记忆和上下文信息, 本节使用 *LSTM* 来处理序列数据. 在注意力机制中使用 *LSTM* 使模型能够更有效地融合不同部分的信息, 从而提高对复杂序列数据的建模能力. 在引入 *LSTM* 后, 使之与 $\widetilde{MHA}$ 形成一个并行结构, 可以省去一个跳跃连接和一个 *Norm* 操作. 这样简化版的 Transformer 表示为:

$$STB(X)=\beta_{FF}LSTM(Norm(X))+\beta_{SA}\widetilde{MHA}(Norm(X)) \tag{12}$$

(2) 使用 *STB* 对单词表示进行编码. 对单词表示向量 S^{w} 进行注意力机制计算的目的是在给定上下文的情况下, 为每个单词赋予不同的注意力权重, 得到任意位置之间的相关性.

首先使用 *Lw* 层 *STB* 对单词表示向量进行加权和向量计算, 对于第 1 层 *STB* 的输出:

$$out_1^{w}=STB_1^{w}(S^{w}) \tag{13}$$

接下对于第 $m\in(2,\cdots,Lw-1)$ 层 *STB* 的输出:

$$out_m=STB_m^{w}\left(out_{m-1}^{w}\right) \tag{14}$$

最终单词加权和向量表示为:

$$S_{\text{out}}^{w}=STB_{Lw}^{w}\left(out_{Lw-1}^{w}\right) \tag{15}$$

其中, $S_{\text{out}}^{w}\in\mathbb{R}^{N\times d}$.

(3) 使用 CNN 来捕捉字符表示 S^{c} 的序列信息, 可以获取到不同位置的局部特征, 有助于识别字符之间的模式和学习字符级别的抽象表示. 字符表示经过 CNN 可以表示为:

$$S_L^{c}=\text{CNN}(S^{c}) \tag{16}$$

接下来使用 *Lc* 层 *STB* 对字符表示进行加权和向量计算. 同时为了融合单词和字符这两种不同粒度的表示, 本文以单词加权和向量 S_{out}^{w} 作为 Query, 然后以 S_L^{c} 作为 Key 和 Value, 即在字符表示中寻找语义相似的部分做 Attention 的融合. 这样本节将式 (9) 替换为:

$$A_h(X)=Softmax\left(\frac{1}{\sqrt{d_k}}S_{\text{out}}^{w}W_h^{Q}(W_h^{K})^{\text{T}}X^{\text{T}}\right) \tag{17}$$

对于第 1 层 *STB* 的输出:

$$out_1^{c}=STB_1^{c}(S^{w}) \tag{18}$$

接下对于第 $m\in(2,\cdots,Lc-1)$ 层 *STB* 的输出:

$$out_m^{c}=STB_m^{c}\left(out_{m-1}^{w}\right) \tag{19}$$

最后单词和字符这两种不同粒度的加权和向量表示为:

$$S_{\text{out}}=STB_{Lc}^{c}\left(out_{Lc-1}^{c}\right) \tag{20}$$

其中, $S_{\text{out}} \in \mathbb{R}^{N\times d}$.

(4) 再次使用STB融合标签概念粒度的向量表示P, 使模型拥有标签级别的概念, 帮助模型理解并找出整个实体, 并将其规范化到同一概念上. 以P作为 Query, 然后以S_{out}作为 Key 和 Value, 即在单词和字符表示中寻找语义相似的部分做注意力融合. 这样将式 (9) 替换为:

$$A_h(X) = Softmax\left(\frac{1}{\sqrt{d_k}} P W_h^Q (W_h^K)^{\mathrm{T}} X^{\mathrm{T}}\right) \tag{21}$$

最终标签概念、单词和字符这 3 种不同粒度的加权和向量表示为:

$$Z = STB(S_{\text{out}}) \tag{22}$$

2.2.4 联合标签解码器

为了解决联合标签集数量庞大的问题, 本文设计了一个新的多层次标签解码器, 如图 4 所示. 联合标签解码器每一层簇的数量都不相同, 并且下一层簇的数量是上一层的整数倍. 为了使得不同层之间的向量可以进行运算, 本文定义了一种向量扩展方式, 即向量$X = (x_1, x_2, \cdots, x_n)$在扩展$k = 3$倍后得到$expand(X, k) = (x_1, x_1, x_1, x_2, x_2, x_2, \cdots, x_n, x_n, x_n)$.

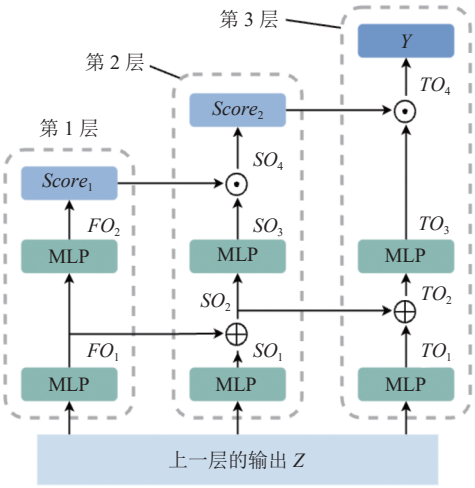

图 4 联合标签解码器

(1) 首先第 1 层, 使用两个多层感知机 (MLP) 来处理纬度高且复杂的非线性关系数据V(记忆模块输出), 可以表示为:

$$\begin{cases} FO_1 = \text{MLP}(Z) \\ FO_2 = \text{MLP}(FO_1) \end{cases} \tag{23}$$

其中, FO_1和FO_2分别表示两次经过 MLP 层后得到的结果.

得到第 1 层所有聚类簇的得分FO_2, 然后将这些簇的得分扩展到原有的K_1倍后得到$Score_1$:

$$Score_1 = expand(FO_2, K_1) \tag{24}$$

(2) 第 2 层. 使用 MLP 处理记忆模块输出Z, 表示为:

$$SO_1 = \text{MLP}(Z) \tag{25}$$

然后与第 1 层的FO_1进行相加融合:

$$SO_2 = SO_1 \oplus FO_1 \tag{26}$$

其中, $\oplus$表示向量相加.

再经过一个 MLP 层:

$$\begin{cases} SO_3 = \text{MLP}(SO_2) \\ SO_4 = SO_3 \odot Score_1 \end{cases} \tag{27}$$

其中, $\odot$表示哈达玛积.

得到第 2 层所有聚类簇的得分, 然后将这些簇的得分扩展到原有的K_2倍后得到$Score_2$:

$$Score_2 = expand(SO_4, K_2) \tag{28}$$

(3) 第 3 层表示为:

$$\begin{cases} TO_1 = \text{MLP}(Z) \\ TO_2 = SO_2 \oplus TO_1 \\ TO_3 = \text{MLP}(TO_2) \\ TO_4 = TO_3 \odot Score_2 \end{cases} \tag{29}$$

最后经过一个 *Softmax* 后得到最终的联合标签预测结果Y:

$$Y = Softmax(TO_4) \tag{30}$$

2.2.5 损失函数

本节模型 MGFFA 使用 BCEWithLogitsLoss 损失函数. 假设有N个 batch, 每个 batch 有n个标签, 则:

$$\text{BCEWithLogitsLoss} = \frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log(\sigma(p_i)) + (1-y_i)\cdot \log(1-\sigma(p_i))) \tag{31}$$

其中, $\sigma(x) = \dfrac{1}{1+\mathrm{e}^{-x}}$是 Sigmoid 函数, log是自然对数.

3 实验与结果分析

3.1 数据集

本文选取了两个广泛应用于生物医学实体识别和关系提取任务的公开生物医学文献数据集: NCBI 疾病语料库和 BC5CDR 数据集.

表 1 提供了两个关键生物医学文献数据集的详细信息概览: NCBI 疾病语料库由美国国家生物技术信息中心 (NCBI) 提供, 包含来自 PubMed 摘要的 793 篇文章, 共标注了 6 881 个疾病实体及其对应的唯一疾病标识符, 覆盖了 1 049 种不同的概念. BC5CDR 数据集是在 BioCreative V 挑战赛中提出的, 包含 1 500 篇 PubMed 文章, 共标注了 28 787 个疾病实体及其对应的唯一疾病标识符, 覆盖了 5 818 种不同的概念.

表 1 BC5CDR 和 NCBI 数据集的相关统计信息

统计值	BC5CDR	NCBI
文章数量	1 500	793
提及数量	28 787	6 881
概念数量	5 818	1 049

3.2 实验设置

本文使用 MGFFA 模型处理 BioNER 和 BioNEN, 输入句长为 50. 记忆矩阵采用 12 头 8 层的注意力机制, 多粒度融合模块结合 8 头 5 层的注意力机制和 2 层 *LSTM*. 联合标签解码器包含 64、512 和 4 096 的聚类数量, 2 层 MLP, 隐藏层为 512, Dropout 为 0.5. 使用 RMSprop 优化器, 初始学习率 1E–5, 批次大小 96, 迭代 80 次.

3.3 评估指标

如果标签为非实体标签的话, 则不进行统计. 在此基础上使用 $F1$ 分数最为评估指标, 表示为:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{32}$$

3.4 基线模型

(1) TaggerOne[28]是一个在训练和预测期间联合 BioNER 和 BioNEN 的机器学习模型, 由半马尔可夫结构化线性分类器组成.

(2) Transiton-based model[29]是一种用于医学领域的疾病命名实体识别和规范化的联合模型, 该模型通过在处理文本时进行状态转换, 动态的识别和标准化疾病名称, 并提高了两个任务的性能.

(3) MTL-MEN&MER_feedback[18]是一个多任务神经网络学习模型, 能够同时处理 BioNER 和 BioNEN 两个任务, 并且通过反馈策略来提升两项任务的性能.

(4) NeuJoRN[17]是一个联合模型, 将 BioNER 和 BioNEN 任务转变为一种动作序列预测任务.

(5) MTAAL[15]是一个用于 BioNER 和 BioNEN 的多任务对抗主动学习模型, 可以有效地选择有助于模型改进的数据样本, 以提高训练效率和模型性能.

(6) E2EMERN[16]是一种端到端的模型, 通过增量任务设置来减少错误传播. 该模型还利用上下文特征来丰富实体提及的语义信息, 以增强 BioNEN 的性能.

3.5 对比实验及其结果分析

MGFFA 模型在 NCBI 和 BC5CDR 数据集上与基线模型对比的实验结果如表 2 所示.

表 2 各模型在 NCBI 和 BC5CDR 数据集上的 $F1$ 分数

Class	Method	NCBI		BC5CDR	
		Recognition	Normalization	Recognition	Normalization
Machine learning	TaggerOne	0.829	0.807	0.826	0.837
	Transition-base	0.821	0.826	0.839	0.856
Deep learning	MTL-MEN&MER_feedback	0.874	0.882	0.876	0.891
	NeubRN	0.886	0.888	0.882	0.899
	MTAAL	0.768	**0.927**	0.860	**0.915**
	E2EMERN	0.915	0.890	0.918	0.897
Ours	MGFFA	**0.930**	0.912	**0.921**	0.910

表 2 显示了 TaggerOne 和 Transition-based 两种机器学习方法在生物医学命名实体识别和规范化任务的性能, 其中 Transition-based 由于使用了非局部特征, 总体表现更佳. 表 2 中还比较了 4 种深度学习方法, 联合模型在两个数据集上普遍优于流水线模型. 其中, MTL-MEN&MER_feedback 在 NCBI 和 BC5CDR 数据集上的 $F1$ 分数分别为 0.874/0.882 和 0.876/0.891, 显示其在 BioNEN 任务上的显著优势. NeuJoRN 在 BioNER 和 BioNEN 任务上的表现至少优于 MTL-MEN&MER_feedback 0.57% 和 0.59%. MTAAL 通过确保任务特定特征的独立性, 并在查询样本多样性方面优化, 提升了性能和主动学习效率. MATTL 在 BC5CDR 的 BioNER 任务中表现最佳, 并在两数据集的 BioNEN 任务上至少优于 NeuJoRN 3.9% 和 1.1%, 在 NCBI 的 BioNEN 任务中达到最高 92.7%. E2EMERN 在 BC5CDR 的 BioNER 任务中也表现突出, 整体优于其他基线模型.

这些基线模型都是通过参数共享将 BioNER 和 BioNEN 任务相关联, 两个任务是松散相关的, 有可能会出现级联错误. 本文提出的 MGFFA 模型使用联合标记方案, 可以避免这种错误的发生, 实现真正的联合输出. MGFFA 模型在 NCBI 数据集和 BC5CDR 数据集的生物医学命名实体识别任务上分别取得了 93% 和 92.1% 的最高 $F1$ 得分.

3.6 消融实验

本文通过消融实验来进一步评估 MGFFA 模型中各个组成部分的贡献, 如表 3 所示.

表 3 消融实验的结果

Method	NCBI		BC5CDR	
	Recognition	Normalization	Recognition	Normalization
MGFFA	0.930	0.912	0.921	0.910
–Memory	0.896	0.884	0.888	0.869
–CF	0.854	0.845	0.867	0.851
–Fusion	0.833	0.795	0.823	0.787
–JTD	0.801	0.792	0.830	0.806

–Memory: 去除整个记忆模块, 使用S_{out}作为概念融合的 Query、Key 和 Value.

–CF: 去除多粒度融合模块的概念融合部分, 使用$Z = S_{\text{out}} \oplus P$作为联合标签解码器的输入.

–Fusion: 去除整个多粒度融合模块, 将预训练词感知表示S^w、预训练字符感知表示S^c和概念表示P相加得到$Z = S^w \oplus S^c \oplus P$, 使用$Z$作为联合标签解码器的输入.

–JTD: 去除联合标签解码器, 使用 *Softmax* 代替.

首先, 去除记忆模块后的结果显示, 模型在 NCBI 数据集上两个任务的 $F1$ 得分分别下降了 3.4% 和 2.8%, 而在 BC5CDR 数据集上下降了 3.3% 和 4.1%, 这表明了记忆模块对于模型的正面影响, 在提升了模型的泛化能力而不会引入额外的复杂度.

其次, 去除多粒度融合模块的概念融合部分的结果显示, NCBI 数据集上生物医学命名实体识别和规范化的 $F1$ 得分分别降低了 7.6% 和 6.7%, BC5CDR 数据集上降低了 5.4% 和 5.9%. 不难看出去除多粒度融合模块的概念融合部分的影响要比去除记忆模块的影响要大. 这是因为概念融合部分在被去除掉后, 记忆模块产生的概念无法被充分利用, 还会对模型产生干扰.

然后尝试移除整个多粒度融合模块. 结果显示, 模型在 NCBI 数据集上的生物医学命名实体识别和规范化任务中的 $F1$ 得分分别下降了 9.7% 和 11.7%, 在 BC5CDR 数据集上下降了 9.8% 和 12.3%. 在未引入多粒度融合信息的情况下, 模型主要依赖于单一粒度的信息处理, 限制了其对于实体边界和内在属性的识别.

最后, 去除联合标签解码器的结果显示, 模型在 NCBI 数据集上的两个任务重的 $F1$ 得分分别下降了 12.9% 和 12%, 在 BC5CDR 数据集上下降了 9.1% 和 10.4%. 联合标签解码器的移除导致模型分别处理每个任务, 从而增加了信息隔阂并降低了处理效率.

综合消融实验的结果得出: MGFFA 模型中的每一个组件都对模型的最终性能有着显著的影响, 这些组件的协同工作使得 MGFFA 模型在生物医学文本的命名实体识别和规范化任务中达到了优异的效果.

4 结论与展望

在本文中, 我们深入探讨了 BioNER 和 BioNEN 任务之间的内在联系, 并提出了一种新颖的多粒度特征融合方法 MGFFA. MGFFA 模型的设计允许我们同时提取 BioNER 和 BioNEN 两个任务的标签, 并通过引入记忆模块形成一种对文本标签的概念级理解, 而多粒度融合模块能够充分利用不同粒度的文本信息来增强模型学习能力. 在 BC5CDR 和 BCBI 数据集上的实验结果验证了 MGFFA 的有效性. 未来, 我们考虑融合文本以外的其他模态数据, 如图像和声音, 来进一步丰富实体识别和规范化的上下文信息.

参考文献

1 Luo L, Wei CH, Lai PT, *et al*. AIONER: All-in-one scheme-based biomedical named entity recognition using deep learning. Bioinformatics, 2023, 39(5): 310–319. [doi: 10.1093/bioinformatics/btad310]

2 Jeon SH, Cho S. Edge weight updating neural network for named entity normalization. Neural Processing Letters, 2023, 55(5): 5597–5618. [doi: 10.1007/s11063-022-11102-2]

3 Dash A, Darshana S, Yadav DK, *et al*. A clinical named entity recognition model using pretrained word embedding and deep neural networks. Decision Analytics Journal, 2024, 10: 100426. [doi: 10.1016/j.dajour.2024.100426]

4 Rani S, Jain A, Kumar A, *et al*. CCheXR-Attention: Clinical concept extraction and chest X-ray reports classification using modified Mogrifier and bidirectional LSTM with multihead attention. International Journal of Imaging Systems and Technology, 2024, 34(1): e23025. [doi: 10.1002/ima.23025]

5 Dai CQ, Zhuang XB, Cai JX. Chinese electronic medical record named entity recognition based on Bi-RNN-LSTM-RNN-CRF. Proceedings of the 11th International Conference on Computing and Pattern Recognition. Beijing: ACM, 2022. 577–583.

6 Guan ZY, Zhou XB. A prefix and attention map discrimination fusion guided attention for biomedical named entity recognition. BMC Bioinformatics, 2023, 24(1): 42. [doi: 10.1186/s12859-023-05172-9]

7 Malmasi S, Fang AJ, Fetahu B, *et al*. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju: International Committee on Computational Linguistics, 2022. 3798–3809.

8 Li JY, Fei H, Liu J, *et al*. Unified named entity recognition as word-word relation classification. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI, 2022. 10965–10973.

9 Toral A, Muñoz R. A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. Proceedings of the 2006 Workshop on NEW TEXT Wikis and Blogs and Other Dynamic Text Sources. 2006. 56–61.

10 Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016. 260–270.

11 Ganea OE, Hofmann T. Deep joint entity disambiguation with local neural attention. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 2619–2629.

12 Mehmood T, Serina I, Lavelli A, *et al*. On the use of knowledge transfer techniques for biomedical named entity recognition. Future Internet, 2023, 15(2): 79–106. [doi: 10.3390/fi15020079]

13 Sung M, Jeon H, Lee J, *et al*. Biomedical entity representations with synonym marginalization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 3641–3650.

14 Peng H, Xiong Y, Xiang Y, *et al*. Biomedical named entity normalization via interaction-based synonym marginalization. Journal of Biomedical Informatics, 2022, 136: 104238. [doi: 10.1016/j.jbi.2022.104238]

15 Zhou BH, Cai XR, Zhang Y, *et al*. MTAAL: Multi-task adversarial active learning for medical named entity recognition and normalization. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI, 2021. 14586–14593.

16 Zhou BH, Cai XR, Zhang Y, *et al*. An end-to-end progressive multi-task learning framework for medical named entity recognition and normalization. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 2021. 6214–6224.

17 Ji ZC, Xia T, Han M, *et al*. A neural transition-based joint model for disease named entity recognition and normalization. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 2021. 2819–2827.

18 Zhao SD, Liu T, Zhao SC, *et al*. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019. 817–824.

19 Yang J, Liang SL, Zhang Y. Design challenges and misconceptions in neural sequence labeling. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: AAAI, 2018. 3879–3889.

20 Li YM, Tao W, Li ZH, *et al*. Artificial intelligence-powered pharmacovigilance: A review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets. Journal of Biomedical Informatics, 2024, 152: 104621. [doi: 10.1016/j.jbi.2024.104621]

21 Xu H, Liu B, Shu L, *et al*. BERT post-training for review reading comprehension and aspect-based sentiment analysis. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 2324–2335.

22 Mayhew S, Nitish G, Roth D. Robust named entity recognition with truecasing pretraining. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 8480–8487.

23 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]

24 Ma WT, Cui YM, Si CL, *et al*. CharBERT: Character-aware pre-trained language model. Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020. 39–50.

25 Sukhbaatar S, Szlam A, Weston J, *et al*. End-to-end memory networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 2440–2448.

26 Graves A, Wayne G, Reynolds M, *et al*. Hybrid computing using a neural network with dynamic external memory. Nature, 2016, 538(7626): 471–476. [doi: 10.1038/nature20101]

27 He B, Hofmann T. Simplifying Transformer blocks. Proceedings of the 12th International Conference on Learning Representations. Vienna, 2024. 1–29.

28 Leaman R, Lu ZY. TaggerOne: Joint named entity recognition and normalization with semi-Markov Models. Bioinformatics, 2016, 32(18): 2839–2846. [doi: 10.1093/bioinformatics/btw343]

29 Lou YX, Zhang Y, Qian T, *et al*. A transition-based joint model for disease named entity recognition and normalization. Bioinformatics, 2017, 33(15): 2363–2371. [doi: 10.1093/bioinformatics/btx172]

(校对责编: 张重毅)