

文章编号: 1006-3080(2025)02-0260-10

DOI: 10.14135/j.cnki.1006-3080.20240426002

一种标签融合驱动的中文医疗实体关系抽取方法

陈 闯, 张维彦, 阮 彤, 郑 红

(华东理工大学信息科学与工程学院, 上海 200237)

摘要: 医疗实体关系抽取是推动医疗信息化建设的关键步骤, 旨在从医疗文本中抽取结构化的三元组信息。针对现有方法对实体类型标签和关系标签利用不充分的问题, 提出了一种标签融合驱动的中文医疗实体关系抽取框架。首先, 将实体关系抽取任务拆分成双向的4个命名实体识别任务, 并将每个任务的标签替换为头尾实体类型标签和关系标签的融合; 其次, 设计了三元组构造策略以最大限度利用双向抽取出的三元组; 最后, 利用三元组双向过滤模型筛选候选三元组。结果表明, 该方法相较于 GPLinker 在 F1 指标上提升了 3.01%。此外, 该方法在医疗领域的重叠关系、多三元组和跨句三元组复杂场景中也表现出了优秀的性能。

关键词: 医疗文本; 实体关系抽取; 标签融合; 双向抽取; 三元组过滤模型

中图分类号: TP391.1

文献标志码: A

实体关系抽取任务旨在从自然语言文本中提取结构化信息, 它有许多下游应用, 例如知识库构建和信息检索等。在生物医学领域, 电子健康记录系统的广泛应用导致涵盖表型描述和患者治疗的医疗文本呈指数级增长^[1]。生物医学实体关系抽取旨在自动识别医学术语并挖掘疾病与药物/症状之间的相互关系^[2]。对于探索临床治疗、病理学、药物发现等研究来说, 这是一项重要的文本挖掘和知识发现任务^[3]。

现有的医疗实体关系抽取方法可根据处理数据的方式和步骤将其划分为管道方法和联合方法。前者将医疗实体关系抽取任务转化为两个子任务, 即实体识别^[4]和关系分类^[5]。首先采用序列标记模型从医疗文本中抽取所有潜在的头实体和尾实体, 然后利用分类器对已抽取的头尾实体对识别它们之间的关系类型, 代表工作有 KGDS^[1], PURE^[6]和 PL-Marker^[7]等。然而, 管道方法忽略了实体和关系之间的额外联合特征, 这导致了两个问题: (1) 错误传播, 因为早期步骤中的错误无法在后续步骤中纠正^[8]; (2) 三元组结构内实体对和关系的相互作用不足^[9]。

为缓解上述问题, 联合方法应运而生。相关工作有 GPLinker^[10], RWG-LSA^[11]等。然而, 这些联合方法专注于各种联合算法, 但忽略了它们本质上基于单独的标签空间^[7]。

与通用领域相比, 医疗领域有更多特定领域的困难。例如: (1) 与通用领域实体关系抽取任务相比, 由于医疗文本独特的文本结构和电子病历专业的记录习惯, 医疗实体关系抽取任务中实体的构成更加复杂^[12], 且存在嵌套实体和重叠关系问题。重叠关系分为实体对重叠(Entity Pair Overlap, EPO)和单实体重叠(Single Entity Overlap, SEO)^[13]。此外, 跨句三元组(头实体和尾实体位于不同句子)和多三元组(一条文本包含多个三元组)复杂场景也给医疗实体关系抽取任务增添了挑战^[2]。(2) 与英文文本相比, 中文文本的实体关系抽取更具有挑战性, 因为中文文本没有明确的单词分隔符^[14]。

为解决上述问题, 本文提出了标签融合驱动的医疗实体关系抽取框架。首先, 将实体关系抽取任务拆分成双向的4个命名实体识别任务, 并将三元组

收稿日期: 2024-04-26

基金项目: 上海市促进产业高质量发展专项资金(2021-GZL-RGZN-01018)

作者简介: 陈 闯(1998—), 男, 内蒙古通辽人, 硕士生, 主要研究方向为自然语言处理。E-mail: 839238634@qq.com

通信联系人: 郑 红, E-mail: zhenghong@ecust.edu.cn

引用本文: 陈 闯, 张维彦, 阮 彤, 等. 一种标签融合驱动的中文医疗实体关系抽取方法[J]. 华东理工大学学报(自然科学版), 2025, 51(2): 260-269.

Citation: CHEN Chuang, ZHANG Weiyan, RUAN Tong, et al. A Chinese Medical Entity Relation Extraction Method Driven by Label Fusion[J]. Journal of East China University of Science and Technology, 2025, 51(2): 260-269.

的头/尾实体标签和关系标签融合,以充分利用标签信息。其次,设计了头/尾实体识别器从两个方向获取头尾实体对。再次,基于关系类型标签进行候选三元组构造。最后,设计了一个过滤模型对候选三元组进行筛选,以获取最终输出。

1 相关工作

医疗实体关系抽取是自然语言处理研究中的一项重要任务,其对众多应用场景有着举足轻重的影响,包括但不限于知识图谱构建、信息检索、问答系统及推荐系统等。现有的医疗实体关系抽取方法主要可分为两类:一是基于管道模型的方法,二是基于联合模型的方法。

管道模型通常将医疗实体关系抽取任务转换为命名实体识别^[4]和关系分类^[5]来作为两个独立的任务。现有的典型方法大致可以分为3类:传统模型、神经网络模型和预训练语言模型。首先,传统模型是基于特征的^[15]方法或基于核函数^[16]的方法,这些方法通常耗时且难以扩展^[8]。其次,神经网络模型聚合来自医学背景的特征,以学习实体和实体对之间的交互信息来进行医学实体识别,然后提取实体对之间的关系类型特征^[17]。此类方法的代表模型包括 KGDS^[1]、BIORE^[18]和 DCNN^[19]。第三,预训练语言模型利用注意力机制在大规模无监督语料库上获得强大的编码能力,可以分别隐式学习更丰富的实体和关系特征。预训练语言模型在医学领域的代表性工作包括 Multi-BERT-wwm^[20]和 KRC^[21]。具体来说,KRC提出了一种用于生物医学实体关系抽取的知识增强阅读理解框架,该框架首先通过询问为每个关系生成问题以识别实体,然后采用基于先验知识的注意机制来指导关系分类。这种方法忽略了实体和关系之间的额外联合特征,这通常会导致模型之间的误差传播。

医疗实体关系抽取任务的联合模型方法是同时识别命名实体并提取它们之间的关系类型^[22],可以分为两种类型,即联合标记方法和参数共享方法。联合标记方法是通过专门设计标记方案,例如级联二进制指针标记框架^[23]和 Handshaking 标记方案^[24],将医疗实体关系抽取任务转化为序列标记任务。特别地,Handshaking 标记方案^[24]对每个关系类型潜在实体对的边界进行标记。参数共享方法在实体识别和关系分类模型中共享特征提取层,其在医学领域的典型工作包括 Span-graph^[25]和 SEGATT-CNN^[22]。具体来说,Span-graph 首先将实体关系抽取任务视为

三元组预测问题,再通过枚举候选实体跨度构建实体图,最后利用跨度评分器和关系评分器捕捉实体间关系,输出三元组。与联合标记方法相比,参数共享方法能够有效地处理更多特征信息,并且更容易实现。

2 任务定义

医疗实体关系抽取任务旨在从句子级医疗文本中抽取实体和关系。给定由 m 个词元(Token)组成的医疗文本 $C = \{c_1, c_2, \dots, c_m\}$,目标是输出文本中由头实体 h 、关系 r 和尾实体 t 组成的所有三元组,即三元组集合 Tri:

$$\text{Tri} = \{(h, r, t) | h, t \in E, r \in R\} \quad (1)$$

其中, $R = \{r_1, r_2, \dots, r_n\}$,表示关系集合,由 n 个预定义的关系类型组成; E 表示实体集合; $H = \{h_1, h_2, \dots, h_l\}$,表示在该文本中有 l 个头实体; $T = \{t_1, t_2, \dots, t_k\}$,表示在该文本中有 k 个尾实体。

3 实验方法

本文提出了一种标签融合驱动的医疗实体关系抽取框架,包含头/尾实体识别器、构造候选三元组和三元组双向过滤模型这3个模块,如图1所示。

3.1 头/尾实体识别器

头/尾实体识别器旨在基于给定医疗文本从两个方向分别获取句子中的头尾实体对,两个方向分别是:先从文本中抽取头实体,再基于抽取的头实体来抽取对应的尾实体,记为 h2rt(Head entity to relation and tail entity);先从文本中抽取尾实体,再基于抽取的尾实体来抽取对应的头实体,记为 t2rh(Tail entity to relation and head entity)。本模块的核心思想是利用基于“实体-融合关系”数据训练的实体识别器来依次抽取头尾实体对,其中融合关系即为头实体类型、关系和尾实体类型的拼接。为了实现这一目的,本文首先重新构造了训练数据,并微调了预训练语言模型(如 BERT^[26])来作为头/尾实体识别器,从两个方向来抽取实体对。由于 h2rt 和 t2rh 的提取过程是相同和独立的,为简单起见,本节仅描述 h2rt 的抽取过程。

3.1.1 训练数据构造 区别于之前的实体关系抽取工作,本文将头/尾实体类型标签融入至关系类型标签上,即将三元组中融合后的关系类型替换头/尾实体的实体类型标签来进行实体识别。原因是:关系标签通常包含了部分实体类型标签的信息,单独建

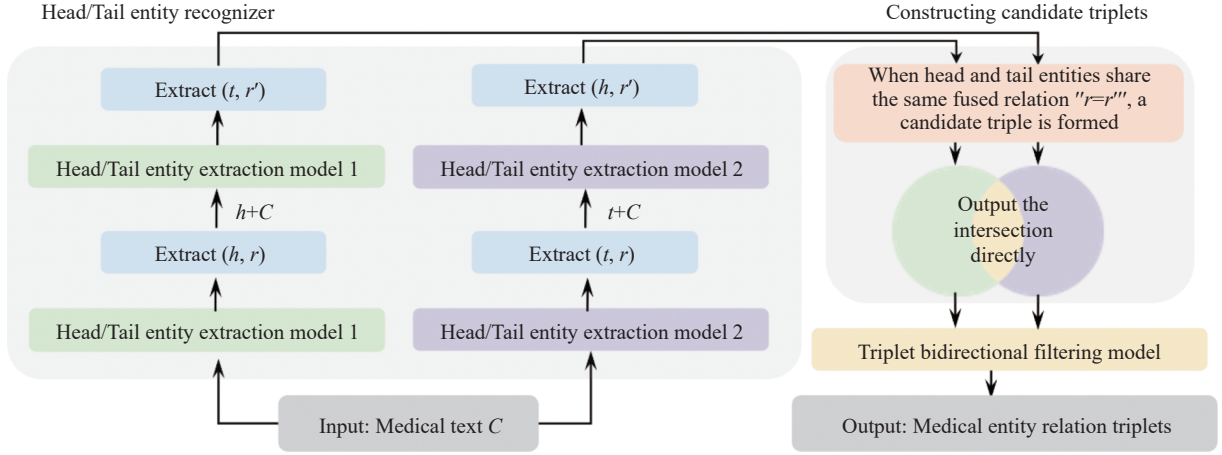


图 1 标签融合驱动的中文医疗实体关系抽取框架

Fig. 1 Chinese medical entity relation extraction framework driven by label fusion

模关系标签和实体类型标签会导致模型重复学习到相同的知识。标签融合有助于充分利用实体标签和类型标签信息,并提高推理效率。具体而言,首先将原始的训练集构造出 4 部分训练数据用于 2 个抽取方向(抽取头实体和根据头实体抽取对应的所有尾实体)。以 h2rt 方向为例,针对第 1 步,将训练数据构造造成“文本:反流样消化不良的突出表现是胃灼热、反流。标签:(头实体:反流样消化不良,融合关系:疾病-临床表现-症状)”。针对第 2 步,将训练数据构造造成“文本:反流样消化不良 [s] 反流样消化不良的突出表现是胃灼热、反流。标签:(尾实体:胃灼热,融合关系:疾病-临床表现-症状),(尾实体:反流,融合关系:疾病-临床表现-症状)”。值得注意的是,同一个抽取方向上的两种数据共用一个模型。

3.1.2 头实体识别器 即根据给定的医疗文本来识别头实体跨度及其对应的融合关系。为了更大地发挥模型能力,将命名实体识别任务转换为实体边界检测任务。在本文中,为了获取头实体跨度 h ,采用标准的基于跨度的模型(如全局指针网络^[27])作为头实体识别器。其核心思想是为每个关系类型分配一个全局矩阵,并记录输入医疗文本中所有头实体跨度的概率。这有效地缓解了医疗实体关系抽取中存在的嵌套实体问题。

形式化地,给定输入医疗文本 C ,首先利用上下文感知的预训练语言模型 PLM(如 BERT^[26])为每个输入词元 c_n 输出文本向量表示:

$$\mathbf{v}_C^H = \text{PLM}([\text{CLS}][c_1, c_2, \dots, c_n][\text{SEP}]) \quad (2)$$

其中, [CLS] 和 [SEP] 是预训练语言模型中的特殊词元。然后,将全局矩阵定义为 $\mathbf{M} \in \mathbb{R}^{n \times n}$,其中 n 为输入医疗文本 C 中的词元个数。矩阵中的每个元素 $M_{s,e} (1 \leq s, e \leq n)$ 表示跨度从第 s 个词元开始到第 e

个词元结束的概率,如式(3):

$$M_{s,e} = (W_1 \mathbf{v}_s + b_1)^T \mathbf{R}_s^T \mathbf{R}_e (W_2 \mathbf{v}_e + b_2) \quad (3)$$

其中, W_1 、 W_2 和 b_1 、 b_2 为可训练权值, $\mathbf{R}_s$ 、 $\mathbf{R}_e$ 为旋转矩阵^[28],分别用于向 $\mathbf{v}_s$ 和 $\mathbf{v}_e$ 中添加相对位置信息。那么,头实体预测器的损失函数如式(4)所示:

$$\mathcal{L}_H^{\text{h2rt}} = \log \left(1 + \sum_{(s,e) \in \Omega_p} e^{-M_{s,e}} \right) + \log \left(1 + \sum_{(s,e) \in (\Omega - \Omega_p)} e^{M_{s,e}} \right) \quad (4)$$

其中, $\Omega = \{(s,e) | 1 \leq s \leq e \leq n\}$ 是给定医疗文本 C 中所有可能的头实体跨度集合, $\Omega_p \subset \Omega$ 包含所有真实的头实体跨度。因此,跨度检测被建模为二分类问题,其中输入为 $M_{s,e}$,输出为真或假,最后从医疗文本中获得头实体及其对应的融合关系类型 (h, r) 。

3.1.3 尾实体识别器 依据已抽取的头实体和对应的医疗文本,以继续获取尾实体及其对应的融合关系。本文设计的尾实体识别器仍然沿用标准的基于跨度的模型。值得注意的是,在 h2rt 方向的抽取过程中,尾实体识别器与前一步头实体识别器的区别仅在于输入不同,因此可以将这两个识别器共同训练。形式化地,给定输入医疗文本 C ,与当前文本中抽取的每个头实体逐一拼接构成多条模型输入,形式记为 $[\text{CLS}]h[s]C[\text{SEP}]$ 。首先利用上下文感知的预训练语言模型为每个输入词元 c_n 输出文本向量表示。如下所示:

$$\mathbf{v}_C^T = \text{PLM}([\text{CLS}]h[s][c_1, c_2, \dots, c_n][\text{SEP}]) \quad (5)$$

其中, $[s]$ 是预训练语言模型中的特殊词元。尾实体识别器对应的损失函数与公式(4)一致,记为 $\mathcal{L}_T^{\text{h2rt}}$ 。最后从医疗文本中获得尾实体及其对应的融合关系类型 (t, r') 。

3.2 构造候选三元组

构造候选三元组模块旨在基于双向抽取的头尾

实体对结果尽可能获取符合事实的候选三元组, 以提高召回和精度。该模块分为两部分: 首先, 针对两个方向的头实体识别器获取的头实体、融合关系类型标签与尾实体识别器获取的尾实体、融合关系类型标签, 将关系类型标签一致的头尾实体构造成候选三元组, 得到两个方向的候选三元组集合。形式化地, 本模块给定同一个方向的头实体识别器结果 (h, r) 和尾实体识别结果 (t, r') , 以 r 和 r' 是否相同为依据来构造候选三元组, 当 r 和 r' 相同时, 则输出候选三元组 (h, r, t) , 如果不同则说明头实体识别和尾实体识别存在分歧, 舍弃该结果。然后, 将两个方向构造的候选三元组取并集作为该模块的最终输出结果, 记为 Tri_c 。最后, 将两个方向的候选三元组集合中的交集部分直接输出作为最终结果, 并集部分则作为最终的候选三元组输入三元组双向过滤模型进一步筛选。

3.3 三元组双向过滤模型

三元组双向过滤模型目标是从候选三元组集合中选择正确的三元组作为最终结果。在此过程中, 对于每个关系类型, 设计了两个全局配对指针网络来分别从给定文本中识别头实体和尾实体的起始/结束标记。如果关系类型在候选三元组集合中的头实体和尾实体的起始/结束位置均与识别的结果一致, 则认为这个三元组是最终的输出。形式化地, 给定医疗文本, 即“[CLS]C[SEP]”, 并将其输入到语言模型来获取文本表示 H_i 。然后, 设计了两个不同的全局配对指针网络。一个网络提取头实体和尾实体的起始标记 (h_s, t_s) , 另一个网络则提取头实体和尾实体的结束标记 (h_e, t_e) 。最后, 使用与关系 r 对应的头尾实体对 (h, t) 来训练上述两个网络, 即

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_e \quad (6)$$

其中, $\mathcal{L}_s$ 和 $\mathcal{L}_e$ 分别是两个全局配对指针网络的损失函数, 参见式(4)。在推理阶段, 对于每个 $(h, r, t) \in \text{Tri}_c$, 分别标记头实体 h 和尾实体 t 的起始和结束词元在给定文本中的位置。因此, 三元组被表示为 (h_s, h_e, r, t_s, t_e) 。然后检查概率 P_{h_s, t_s} 和 P_{h_e, t_e} , 它们对应于全局配对指针网络矩阵中的位置 (h_s, t_s) 和 (h_e, t_e) 。如果这些概率分别超过预定义的阈值 λ_s 和 λ_e , 将保留三元组作为最终输出的三元组。

4 实验部分

4.1 实验设置

4.1.1 数据集 为了评估本文提出的方法, 本文使用

中文医疗实体关系抽取数据集 CMeIE-V2 进行实验, 其来自中文生物医学语言理解评估 2.0 [29]。CMeIE-V2 的详细统计结果如表 1 所示。由于 CMeIE-V2 测试标签尚未发布, 无法根据三元组情况来统计测试集。因此, 复杂场景的分析实验会在验证集上进行。

表 1 CMeIE-V2 数据集统计

Table 1 CMeIE-V2 dataset statistics

Class	Train	Valid	Test
Entity	11	11	11
Relation	44	43	44
Sample	14339	3585	4482
Entity overlap type (Normal)	1910	444	—
Entity overlap type (EPO)	58	19	—
Entity overlap type (SEO)	12371	3122	—
Triplets in a sample ($N = 1$)	1722	399	—
Triplets in a sample ($N = 2$)	2588	672	—
Triplets in a sample ($N = 3$)	2393	600	—
Triplets in a sample ($N = 4$)	2003	487	—
Triplets in a sample ($N \geq 5$)	5567	1409	—
Triplets distance (Single-sentence)	10098	2636	—
Triplets distance (Cross-sentence)	4241	949	—

4.1.2 对比基线 在中文医疗数据集 CMeIE-V2 上, 本文将所提方法与两类基线方法进行比较: 一类为管道方法, 即 CBLUE[29] 官方给出的方法, 并使用了多个预训练语言模型进行对比。另一类为联合方法, 包括 CasRel[30]、Cas_CLN&BTCAMS-mhs [23] 和 GPLinker[10], 预训练语言模型均使用 RoBERTa_zh_Large_PyTorch。在英文通用数据集 ACE05 上, 本文将与 TableSeq[31]、ATG[32]、PURE[6]、PL-Marker[7] 和 recollect-tuning[33] 进行比较, 其中, ATG 的预训练语言模型使用预训练的 transformer[34] 的编码器和随机初始化的解码器, 其余方法均使用 ALBERT_{XXLARGE}。

4.1.3 评估指标 对于数据集 CMeIE-V2, 本文采用精确度 (P)、召回率 (R) 和微平均 $F1$ 分数 ($F1$) 进行评估。由于 CMeIE-V2 测试标签未发布, 最终评估分数是通过将整个预测文件提交至官网获得的, 需要 (h, r, t) 结果准确匹配。对于数据集 ACE05, 本文采用 Ent, Rel 和 Rel+这 3 个指标表示不同任务的 $F1$ 。其中, Ent 表示实体预测, 将头实体和尾实体都视为实体, 当且仅当实体和实体类型均正确时认为实体预测是正确的; Rel 表示关系预测, 当 (h, r, t) 均正确时认为关系预测是正确的; Rel+表示严格关系预测,

还需要准确的实体类型预测。

4.1.4 参数设定 硬件方面,使用 Intel(R) Xeon(R) Gold 6139M CPU、128 GB 内存及 4 块 NVIDIA GeForce RTX 3090 24G GPU。模型超参数中,输出序列的最长 token 数为 256,头/尾实体识别器的 batch_size 设为 16,总共训练 20 个 epoch,学习率设为 4×10^{-5} ,三元组双向过滤模型的 batch_size 设为 8,总共训练 40 个 epoch,学习率设为 2×10^{-5} 。为了保证实验的可复现性,设置随机种子为 2021。

4.2 主实验结果

首先,本文在数据集 CMeIE-V2 和 ACE05 上进行了对比实验。CMeIE-V2 上的实验结果如表 2 所示,可以得出如下结论:(1)本文所提方法在 $F1$ 上优于所有的对比基线实验,证明了该方法的有效性。在 CMeIE-V2 上,本方法达到了该任务的最佳性能 57.27%。(2)在所有基线中, RoBERTa-wwm-ext-large 在管道方法中表现最好,但本方法相比管道方法在 $F1$ 上仍然高出 7.04%。原因可能是在管道式框架下,将实体类型标签融合至关系标签上,实现标签信息融合有助于提高标签信息的利用率。(3)将本方法与联合方法的 GPLinker 进行比较,本方法在 $F1$ 上仍然提升 3.01%,这表明了子任务融合训练策略的有效性,专注于保持训练与推理任务的一致性,可以充分挖掘模型潜在的最优性能,即使是管道式的抽取框架依旧能优于联合模型。

ACE05 上主实验结果如表 3 所示,由实验结果可知,本方法在 Rel 和 Rel+指标上均为最优,相比次优模型 Recollect-tuning 在 Rel+指标上提升了 0.8%,这表明了本方法在实体关系抽取任务上的有效性。

4.3 细节分析实验

4.3.1 消融实验 为了评估核心组件(每个创新设计点及其核心模块)的影响,对本方法进行了消融分析,实验结果详见表 4。其中, A 表示混合训练策略, B 表示实体信息增强方法, filter 表示三元组双向过滤模型, Ours-filter($\cup$) 和 Ours-filter($\cap$) 分别表示将双向抽取的结果进行并集和交集操作。h2rt 和 t2rh 表示双向抽取的两个独立抽取方向, T 表示三元组构造策略。然后,依次删除各个组件并评估对剩余组合的性能。

从表 4 可以得出:(1)删除任何组件都会导致本方法性能的下降。具体来说,删除过滤模型带来的性能下降是最大的,导致本方法在 $F1$ 指标上直接下降 2.87%。这证明了本方法所设计的三元组双向过滤模型是十分有效的,基于识别并构造大量候选三元组保证高召回,随后通过设计过滤模型来删除大

表 2 数据集 CMeIE-V2 上的主实验

Method	Baseline	P/%	R/%	F1/%
Pipeline method	ALBERT-tiny	39.12	35.06	36.98
	BERT-base	48.81	47.26	48.02
	ZEN-base	49.96	46.25	48.03
	MedBERT-base	48.84	47.66	48.24
	PCL-MedBERT	49.34	47.66	48.49
	MacBERT-base	49.17	47.99	48.57
	BERT-wwm-ext	49.66	47.61	48.61
	RoBERTa-wwm-ext-base	49.34	48.32	48.82
	ALBERT-xxlarge	51.20	47.85	49.47
	MacBERT-large	51.35	48.50	49.89
	RoBERTa-large	51.55	48.77	50.12
	RoBERTa-wwm-ext-large	51.71	48.84	50.23
	Ours	58.98	55.66	57.27
Joint method	CasRel	51.86	33.82	40.94
	Cas_CLN&BTCAMS-mhs	53.16	48.61	50.79
	GPLinker	56.34	52.33	54.26

表 3 数据集 ACE05 上的主实验

Method	Ent/%	Rel/%	Rel+/%
TableSeq	89.5	67.6	64.3
ATG	90.1	68.7	66.2
PURE	90.9	69.4	67.0
PL-Marker	91.1	73.0	71.1
Recollect-tuning	91.5	73.6	71.5
Ours	84.0	73.6	72.3

部分错误三元组保证高精度。(2)与 Ours-h2rt 和 Ours-t2rh 方法对比,本方法在 $F1$ 上分别下降 1.74% 和 1.44%。可以发现在 CMeIE-V2 数据集中,从头实体到尾实体的抽取方向比从尾实体到头实体的抽取方向发挥着更重要的作用。这可能是由于 CMeIE-V2 数据集中头实体和尾实体类型分布不平衡,头实体类型大多为“疾病”,尾实体类型分布均匀,从而使同一个头实体可能与更多的尾实体构成关系,使得 h2rt 方向的模型训练时在全局矩阵中标记为真的个数更多,模型训练更充分。因此,在这种情况下,本方法所采用的双向抽取是十分有效的。(3)对比 Ours-AB, Ours-A, Ours-B 实验,虽然混合训练策略和

实体信息增强方法叠加带来的增益不及单一模块增益加和的大,但也使 $F1$ 提升 1.68%,说明混合训练策略与实体信息增强方法对本方法有不可替代的作用。(4)删除构造候选三元组模块后, $F1$ 精度下降 0.34%,证明了本文提出的候选三元组的构造策略是有效的。

表 4 消融实验

Table 4 Ablation experiment			
Method	P/%	R/%	F1/%
Ours-A	61.95	50.79	55.82
Ours-B	58.93	53.32	55.99
Ours-AB	60.02	51.77	55.59
Ours-filter($\cup$)	52.00	59.54	55.51
Ours-filter($\cap$)	65.77	46.38	54.40
Ours-h2rt	61.35	50.71	55.53
Ours-t2rh	61.37	51.21	55.83
Ours-T	59.44	54.63	56.93
Ours	58.98	55.66	57.27

4.3.2 子任务融合训练策略分析实验 为了分析不同子任务数据融合训练对模型性能的影响,针对本文提出的标签驱动的中文医疗实体关系抽取框架,设计了 5 种融合训练策略。方案信息如表 5 所示,其中, $[s1] \sim [s6]$ 表示分隔标记。具体来说,针对每一种融合训练策略,本文分别罗列了需要训练的模型

数,每个模型对应训练数据结构及对应的训练数据条数,并标注了所训练的模型该在哪个方向上使用。

为了验证上述融合训练策略之间的性能差异,本文基于所提框架在 CMeIE-V2 测试集上进行了大量实验。实验结果如表 6 所示,可以得出结论:(1)在 5 种混合训练策略中,策略 5 的效果均高于其他策略,两个抽取方向上的 $F1$ 分别达到 55.47% 和 54.72%。这可能是因为策略 5 不仅保持了训练任务和推理任务数据结构高度一致且数量相同,而且抽取头实体和根据头实体抽取尾实体、抽取尾实体和根据尾实体抽取头实体这两对子任务间的共有特征更利于模型训练。(2)对于混合训练策略 1,用所有的数据训练同一个模型,但模型效果却是所有策略中最差的。这可能是因为当任务和数据类别增加时,类数据特征学习不充分,反而造成性能下降。这也从侧面说明了联合模型存在的弊端。

4.3.3 实体信息增强分析实验 针对根据头实体抽取尾实体以及根据尾实体抽取头实体这种实体配对任务,本节设计了多种方法以提示模型当前配对任务的前置实体的相关信息。具体来说,设计了 3 种提示模板前缀和 3 种嵌入手段,组合实验方案如表 7 所示。其中,嵌入方案中, None 表示不使用嵌入, Average 和 Attention^[34] 分别表示将头实体头尾位置的编码后的嵌入表示直接求平均和进行注意力计算后求平均,再加入到每一个词上。前缀方案中, ETRC(Entity entity type relation context)表示将实体、实体类型和关系拼接到文本前, ERC(Entity relation context)表示

表 5 子任务融合训练策略

Table 5 Subtask fusion training strategy					
Training strategy	Model count	Training data structure	Sample	Direction 1 (h2rt)	Direction 2 (t2rh)
1	1	Model 1: $[s1][s2][s3] c \rightarrow (h, r)$, $[s4][s5][s6] c \rightarrow (t, r)$, $h[s1]c \rightarrow (t, r), t[s1]c \rightarrow (h, r)$	114 100	Model 1	Model 1
2	2	Model 2: $[s1][s2][s3] c \rightarrow (h, r)$, $[s4][s5][s6] c \rightarrow (t, r)$	28 678	Model 2, Model 3	Model 2, Model 3
		Model 3: $h[s1]c \rightarrow (t, r), t[s1]c \rightarrow (h, r)$	85 422		
		Model 4: $c \rightarrow (h, r)$	14 339	Model 4, Model 6	Model 5, Model 7
3	4	Model 5: $c \rightarrow (t, r)$	14 339		
		Model 6: $h[s1]c \rightarrow (t, r)$	22 970		
		Model 7: $t[s1]c \rightarrow (h, r)$	62 452	Model 8, Model 9	Model 9, Model 8
4	2	Model 8: $c \rightarrow (h, r), t[s1]c \rightarrow (h, r)$	76 791		
		Model 9: $c \rightarrow (t, r), h[s1]c \rightarrow (t, r)$	37 309	Model 10	Model 11
		Model 10: $c \rightarrow (h, r), h[s1]c \rightarrow (t, r)$	37 309		
5	2	Model 11: $c \rightarrow (t, r), t[s1]c \rightarrow (h, r)$	76 791		

表 6 子任务融合训练策略分析实验
Table 6 Subtask fusion training strategy analysis experiment

Training strategy	Direction	P/%	R/%	F1/%
1	h2rt	58.36	42.32	49.07
	t2rh	55.09	49.34	52.06
2	h2rt	60.39	42.15	49.64
	t2rh	56.60	49.88	53.03
3	h2rt	59.54	50.50	54.65
	t2rh	56.48	53.05	54.71
4	h2rt	60.48	50.42	54.99
	t2rh	57.42	51.85	54.49
5	h2rt	58.73	52.55	55.47
	t2rh	55.91	53.59	54.72

将实体和关系拼接到文本前, EC(Entity context)表示只将实体拼接到文本前。表中的 h_type 和 t_type 分别表示头实体类型和尾实体类型。有两点值得注意: (1)不同的前缀提示模版对应的训练数据构造不同; (2)由于 ETRC 和 ERC 的前缀包含了本文所提融

合标签中的部分信息, 因此为避免信息泄漏, 这两种方法的实体抽取的标签采用实体类型, EC 方法则使用本文所提的融合标签。

为验证上述实体信息增强方法之间的性能差异, 本节在 CMeIE-V2 测试集上进行了实验。实验结果详见表 7。从表 7 中可以得出: (1)对于 3 种提示模版前缀, EC 方法的效果最好。这种直接拼接头/尾实体的前缀要比其他两种前缀大约提升约 1%~3%。这有效地说明了标签融合带来的巨大提升。(2)对于 3 种嵌入方案, 可以发现 Attention 方式对于 EC 方法的增益相对最大。基于此, 本文所提方法选择 Attention 嵌入组合 EC 前缀的方案。

4.4 复杂场景分析实验

中文医疗领域的专业性和特殊性使得医疗文本包含的信息量更多, 存在大量复杂场景, 如重叠关系、多三元组和跨句三元组, 本文针对这 3 个场景进行分析实验。由于 CMeIE-V2 测试标签尚未发布, 本节实验均在验证集上进行, 且采用 F1 作为评价指标。

4.4.1 重叠关系场景分析实验 由实验结果(表 8)可知, 本文所提出的方法优于无实体重叠和单实体重叠模式中的所有基线, 尤其在单实体重叠上相比

表 7 实体信息增强分析实验
Table 7 Entity information enhancement analysis experiment

Embedding	Type	Prefix	Direction	P/%	R/%	F1/%
None	ETRC	$h[s1]h_type[s2]r[s3]$	h2rt	44.58	56.13	49.69
	ETRC	$t[s1]t_type[s2]r[s3]$	t2rh	50.84	54.09	52.42
	ERC	$h[s1]r[s2]$	h2rt	53.19	55.39	54.27
	ERC	$t[s1]r[s2]$	t2rh	52.94	53.97	53.45
	EC	$h[s1]$	h2rt	56.71	53.91	55.27
	EC	$t[s1]$	t2rh	54.63	53.95	54.29
Average	ETRC	$h[s1]h_type[s2]r[s3]$	h2rt	44.67	55.26	49.40
	ETRC	$t[s1]t_type[s2]r[s3]$	t2rh	50.45	54.01	52.17
	ERC	$h[s1]r[s2]$	h2rt	44.94	56.19	49.94
	ERC	$t[s1]r[s2]$	t2rh	49.66	53.77	51.63
	EC	$h[s1]$	h2rt	57.52	52.94	55.13
	EC	$t[s1]$	t2rh	55.27	53.79	54.52
Attention	ETRC	$h[s1]h_type[s2]r[s3]$	h2rt	45.07	54.92	49.51
	ETRC	$t[s1]t_type[s2]r[s3]$	t2rh	50.81	53.58	52.16
	ERC	$h[s1]r[s2]$	h2rt	45.76	56.10	50.40
	ERC	$t[s1]r[s2]$	t2rh	50.43	53.42	51.88
	EC	$h[s1]$	h2rt	58.73	52.55	55.47
	EC	$t[s1]$	t2rh	55.91	53.59	54.72

GPLinker 在 $F1$ 精度上提高了 2.1%。这可能是由于本方法在子任务的拆分和模型设计上解决了重叠关系问题,并且在头/尾实体对抽取阶段进行双向抽取时保证高召回,然后通过三元组过滤模型来保证高精度。

4.4.2 多三元组场景分析实验 在多三元组场景中,样本包含的三元组数量越多,包含 SEO 和 EPO 模式的概率就越高,这给提取方法带来了多重挑战。表 9 表明本方法在所有三元组数量(N)下均优于基线,且随着三元组数量的增加,本方法相比 GPLinker 提升的效果越来越明显,这在一定程度上表明了指针网络更适合多三元组场景。当 $N \geq 5$ 时,本方法相比 GPLinker, $F1$ 提升了 2.4%。这可能是由于本方法双向抽取和过滤模型的设计使得模型可以有效挖掘文本中潜在的三元组。

4.4.3 跨句三元组场景分析实验 表 10 示出了跨句

表 8 重叠关系场景分析实验

Table 8 Overlapping relation scenarios analysis experiment

Method	Normal	EPO	SEO
CasRel	38.3	35.1	40.9
RoBERTa	43.0	48.9	49.3
Cas_CLN&BTCAMS-mhs	43.5	52.3	50.7
GPLinker	47.4	54.2	53.5
Ours	48.8	54.2	55.6

表 9 多三元组场景分析实验

Table 9 Multiple triplets scenarios analysis experiment

Method	Multiple triplets in a sample				
	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N \geq 5$
CasRel	44.9	42.6	40.9	39.6	40.7
RoBERTa	48.7	47.0	48.7	46.9	50.1
Cas_CLN&BTCAMS-mhs	48.6	51.0	49.9	48.9	50.9
GPLinker	51.9	53.8	52.3	50.5	54.1
Ours	52.0	54.4	53.5	52.5	56.5

表 10 跨句三元组场景分析实验

Table 10 Cross-sentence triplets scenarios analysis experiment

Method	Single-sentence	Cross-sentence
CasRel	41.2	40.1
RoBERTa	49.5	48.5
Cas_CLN&BTCAMS-mhs	50.4	50.4
GPLinker	53.2	53.4
Ours	55.5	55.1

三元组场景分析实验结果。可以看出,本文方法在单句和跨句场景上的性能相比 GPLinker 分别提升了 2.3% 和 1.7%,表明本文方法在一定程度上减少了提取过程中实体距离造成的影响。

5 结束语

本文重点研究了中文医疗实体关系抽取任务中现有方法没有充分利用三元组内头/尾实体标签和关系标签的问题。为了解决该问题,本文在标签融合的基础上提出了包括头/尾实体识别器、候选三元组构造和三元组双向过滤模型 3 个关键模块的中文医疗实体关系抽取框架。此外,还提出了一种子任务融合训练策略并针对实体信息增强方法进行了对比实验,以充分挖掘模型潜在的最优性能。实验证明本文方法性能超过了其他基线方法。同时本文方法即使在涉及重叠关系、多个三元组和跨句三元组复杂场景中也能实现出色的性能增益。未来的工作包括进一步优化框架的效率以及将标签融合驱动的思想拓展到事件抽取或更复杂的文本结构化任务上。

参考文献:

- [1] ZHAO Q, XU D, LI J, *et al.* Knowledge guided distance supervision for biomedical relation extraction in Chinese electronic medical records[J]. *Expert Systems with Applications*, 2022, 204: 117606.
- [2] ZHAO D, WANG J, LIN H, *et al.* Biomedical cross-sentence relation extraction via multihead attention and graph convolutional networks[J]. *Applied Soft Computing*, 2021, 104: 107230.
- [3] LUO R, SUN L, XIA Y, *et al.* BioGPT: Generative pre-trained transformer for biomedical text generation and mining[J]. *Briefings in Bioinformatics*, 2022, 23(6): bbac409.
- [4] AN Y, XIA X, CHEN X, *et al.* Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF[J]. *Artificial Intelligence in Medicine*, 2022, 127: 102282.
- [5] HE B, GUAN Y, DAI R. Convolutional gated recurrent units for medical relation classification[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid, Spain: IEEE, 2018: 646-650.
- [6] ZHONG Z, CHEN D. A Frustratingly easy approach for entity and relation extraction[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 50-61.

- [7] YE D, LIN Y, LI P, *et al.* Packed levitated marker for entity and relation extraction[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: Association for Computational Linguistics, 2022: 4904-4917.
- [8] LI W, ZHANG P, WEI F, *et al.* A novel feature-based approach to Chinese entity relation extraction[C]//Proceedings of ACL-08: HLT. Columbus, OH, USA: Association for Computational Linguistics, 2008: 89-92.
- [9] SHANG Y M, HUANG H, MAO X. OneRel: Joint entity and relation extraction with one module in one step[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(10): 11285-11293.
- [10] 苏剑林. GPLinker: 基于 GlobalPointer 的实体关系联合抽取[J]. *计算机技术与发展*, 2024, 34(3): 147-154.
- [11] 张利, 张欢欢, 袁玉波. 中文重叠关系抽取的动态分层级联标记模型[J]. *华东理工大学学报(自然科学版)*, 2024, 50(3): 450-458.
- [12] 郭瑞, 张欢欢. 基于 RoBERTa 和对抗训练的中文医疗命名实体识别[J]. *华东理工大学学报(自然科学版)*, 2023, 49(1): 144-152.
- [13] LUO L, LAI P T, WEI C H, *et al.* BioRED: A rich biomedical relation extraction dataset[J]. *Briefings in Bioinformatics*, 2022, 23(5): bbac282.
- [14] CAI T, MA Z, ZHENG H, *et al.* NE-LP: Normalized entropy-and loss prediction-based sampling for active learning in Chinese word segmentation on EHRs[J]. *Neural Computing and Applications*, 2021, 33(19): 12535-12549.
- [15] GU J, QIAN L, ZHOU G. Chemical-induced disease relation extraction with various linguistic features[J]. *Database*, 2016, 2016: baw042.
- [16] ZHANG M, ZHOU G, AW A. Exploring syntactic structured features over parse trees for relation extraction using kernel methods[J]. *Information Processing & Management*, 2008, 44(2): 687-701.
- [17] LI Z, YANG J, GOU X, *et al.* Recurrent neural networks with segment attention and entity description for relation extraction from clinical texts[J]. *Artificial Intelligence in Medicine*, 2019, 97: 9-18.
- [18] SOUSA D, COUTO F M. Biomedical relation extraction with knowledge graph-based recommendations[J]. *IEEE Journal of Biomedical and Health Informatics*, 2022, 26(8): 4207-4217.
- [19] SUN X, MA L, DU X, *et al.* Deep convolution neural networks for drug-drug interaction extraction[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid, Spain: IEEE, 2018: 1662-1668.
- [20] ZHANG Y, LIAO X, CHEN L, *et al.* Multi-bert-wwm model based on probabilistic graph strategy for relation extraction[C]//Health Information Science: 10th International Conference, HIS 2021. Melbourne, VIC, Australia: Springer, Cham, 2021: 95-103.
- [21] CHEN J, HU B, PENG W, *et al.* Biomedical relation extraction via knowledge-enhanced reading comprehension[J]. *BMC Bioinformatics*, 2022, 23(1): 20.
- [22] BAI T, GUAN H, WANG S, *et al.* Traditional Chinese medicine entity relation extraction based on CNN with segment attention[J]. *Neural Computing and Applications*, 2022, 34(4): 1-10.
- [23] CHANG H, ZAN H, GUAN T, *et al.* Application of cascade binary pointer tagging in joint entity and relation extraction of Chinese medical text[J]. *Mathematical Bioscience and Engineering*, 2022, 19(10): 10656-10672.
- [24] YANG N, PUN S H, VAI M I, *et al.* A unified knowledge extraction method based on BERT and handshaking tagging scheme[J]. *Applied Sciences*, 2022, 12(13): 6543.
- [25] FEI H, ZHANG Y, REN Y, *et al.* A span-graph neural model for overlapping entity relation extraction in biomedical texts[J]. *Bioinformatics*, 2021, 37(11): 1581-1589.
- [26] DEVLIN J, CHANG M-W, LEE K, *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding[C]//North American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.
- [27] SU J, MURTADHA A, PAN S, *et al.* Global pointer: Novel efficient span-based approach for named entity recognition[EB/OL]. (2022-8-5)[2022-12-11]. <https://doi.org/10.48550/arXiv.2208.03054>.
- [28] SU J, AHMED M, LU Y, *et al.* Roformer: Enhanced transformer with rotary position embedding[J]. *Neurocomputing*, 2024, 568: 127063.
- [29] ZHANG N, CHEN M, BI Z, *et al.* CBLUE: A Chinese biomedical language understanding evaluation benchmark [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: Association for Computational Linguistics, 2022: 7888-7915.
- [30] WEI Z, SU J, WANG Y, *et al.* A novel cascade binary tagging framework for relational triple extraction[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 1476-1488.
- [31] WANG J, LU W. Two are better than one: Joint entity and relation extraction with table-sequence encoders[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 1706-1721.
- [32] ZARATIANA U, TOMEH N, HOLAT P, *et al.* An autoregressive text-to-graph framework for joint entity and relation extraction[J]. *Proceedings of the AAAI Conference on*

- Artificial Intelligence, 2024, 38(17): 19477-19487.
- [33] WU Y, CHEN Y, QIN Y, *et al.* A recollect-tuning method for entity and relation extraction[J]. *Expert Systems with Applications*, 2024, 245: 123000.
- [34] VASWANI A, SHAZEER N, PARMAR N, *et al.* Attention is all you need[J]. *Advances in Neural Information Processing Systems*, 2017, 30: 1-10.

A Chinese Medical Entity Relation Extraction Method Driven by Label Fusion

CHEN Chuang, ZHANG Weiyan, RUAN Tong, ZHENG Hong
(School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: The extraction of medical entity relationships is an key step in promoting medical informationization construction aiming at extracting structured triplet information from medical text. A label fusion-driven Chinese medical entity relation extraction framework is proposed to address the issue of insufficient utilization of entity type labels and relation labels in existing methods. Firstly, the entity relationship extraction task is split into four bidirectional named entity recognition tasks, and the labels of each task are replaced with a fusion of head and tail entity type labels and relation labels. Secondly, a triplet construction strategy is designed to maximize the utilization of the triplets extracted bidirectionally. Finally, a triplet bidirectional filtering model is utilized to filter the candidate triplets. The experimental results show that this method has improved the $F1$ index by 3.01% compared to GPLinker. In addition, this method has also demonstrated excellent performance in complex scenarios such as overlapping relationships, multiple triplets, and cross-sentence triplets in the medical field.

Key words: medical text; entity relation extraction; label fusion; bidirectional extraction; triplet filter model

(责任编辑: 李娟)