doi：10. 12068/j. issn. 1005－3026. 2026. 20240232

基于结构化梯度树提升的知识图谱集体实体消歧方法

刘 军，朱鸿兆，张 昭
（东北大学 计算机科学与工程学院，辽宁 沈阳 110169）

摘 要：为提升医学决策支持知识图谱的全面性和精确性，针对医学领域中实体文本长、复杂且专业性高，以及同一实体在不同数据库中的歧义问题，提出一种基于结构化梯度树提升的集体消歧方法（CED-SGTB）. 首先，清洗和处理医学决策文本，利用神经网络协作检测和标注文本中的实体. 其次，通过全局特征优化当前候选实体和先前实体，完成实体的全局消歧. 再次，设计黄金路径双向集束搜索算法，以减少模型方差. 最后，通过对比实验和消融实验，结果表明CED-SGTB在精确度和调和平均值*F*1方面优于传统方法，能够更精确地完成实体消歧任务.

关 键 词：医学决策支持；知识图谱；梯度树提升；集束搜索；实体消歧
中图分类号：TP 182 **文献标志码：**A **文章编号：**1005-3026(2026)03-0010-10

Collective Entity Disambiguation Method for Knowledge Graph Based on Structured Gradient Tree Boosting

LIU Jun, *ZHU Hong-zhao*, *ZHANG Zhao*
（School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: ZHU Hong-zhao, E-mail: 1024921337@qq.com）

Abstract: To enhance the comprehensiveness and accuracy of the medical decision support knowledge graph, a collective entity disambiguation method based on structured gradient tree boosting (CED-SGTB) was proposed to address challenges in the medical domain, including long, complex, and highly specialized entity texts, as well as the ambiguity of the same entity across different databases. Firstly, medical decision texts were cleaned and processed, and neural networks were deployed to collaboratively detect and annotate entities within the texts. Secondly, global entity disambiguation was achieved by optimizing the current candidate entities and previous entities through global features. Thirdly, a golden path bidirectional beam search algorithm was designed to reduce model variance. Finally, the results of comparative experiments and ablation experiments demonstrate that CED-SGTB outperforms traditional methods in terms of accuracy and F1-score and can more accurately complete entity disambiguation tasks.

Key words: medical decision support; knowledge graph; gradient tree boosting; beam search; entity disambiguation

医学决策支持[1]是利用工具和方法，结合患者健康数据和医学知识，帮助医疗团队作出临床决策. 通过精准评估和实时监控病情，帮助医生更有效地掌握患者状况，从而作出可靠的治疗决策.

然而，医学决策支持也面临诸多复杂挑战，可能受到患者个体差异、突发疾病、环境因素等多方面影响，各类决策之间也存在相互关联与制约. 此外，医疗数据来源丰富、类型多样，对实时性和准确性要求极高且决策价值巨大，这使得大

收稿日期：2024-12-19
基金项目：国家自然科学基金资助项目（61701100）.
作者简介：刘 军（1969—），男，辽宁沈阳人，东北大学教授.
通信作者：朱鸿兆，E-mail：1024921337@qq.com.

量临床数据无法被准确获取和及时处理,从而导致严重的医疗后果[2].因此,高效处理医学决策的复杂信息,全面建立更为完善的医学决策管理系统已成为必然趋势.

现有研究表明,知识图谱技术[3]在医疗领域的应用主要集中在知识表示和融合方面.虽然基于上下文信息的实体链接技术提升了准确率,但对于长尾实体的处理仍有不足.

结构化梯度方法[4]在医疗数据分析中展现出优势,特别是在症状-疾病关联预测等任务中表现出优异的特征学习和泛化能力.然而,这些方法在专业术语理解和计算效率方面仍有提升空间.最新的实体消歧技术结合了深度学习与图神经网络,通过端到端学习自动提取深层语义特征,但面临标注数据需求大、模型复杂度高等应用限制.

实体消歧[5]技术直接关系到知识库的构建质量.当前主流方法主要分为3类:基于规则的传统方法在处理规范化医疗文本时效果稳定但泛化性差;基于统计学习的消歧算法通过特征工程提升了适应性,却难以捕捉专业术语的深层语义;近年来兴起的神经网络方法虽然在准确率上有明显突破,但对标注数据的依赖严重制约了其在临床场景中的应用.由于医疗实体常存在一名多义和一义多名的双重歧义特征,这对消歧算法提出了更高要求[6].

面对全面的医学决策支持系统建设需求[7],本文提出将知识图谱应用于医学领域,重点研究知识图谱构建过程中的知识融合部分.通过探索集体实体消歧的方法,为知识图谱的构建提供有力支持,提升其完备性、丰富性和稳定性,从而增强医学决策支持系统的效能,最终形成完善的医学知识库[8].

1　问题描述及总体研究思路

由于医学领域的决策支持信息具有一定专业性,且实体间存在重要关联特征,以往的方法只能局部优化提及实体,且搜索速度欠佳[9].

本文针对以上问题,提出CED-SGTB方法进行集体实体消歧.首先将获取的医学决策文本进行清洗和处理,得到更规范的数据输入;通过多个神经网络协作对文本中的提及实体进行检测和标注,信息补充后采用基于规则的候选实体生成方法完成实体生成任务;使用当前实体和已分配实体的整个决策历史定义的全局特征,完成对所有提及实体的全局优化;在训练推理过程中,为解决归一化困难的问题,本文对传统集束搜索进行改进,设计黄金路径双向集束搜索(BiBSG)算法,降低模型方差,提高消歧的准确率和效率.总体方案如图1所示.

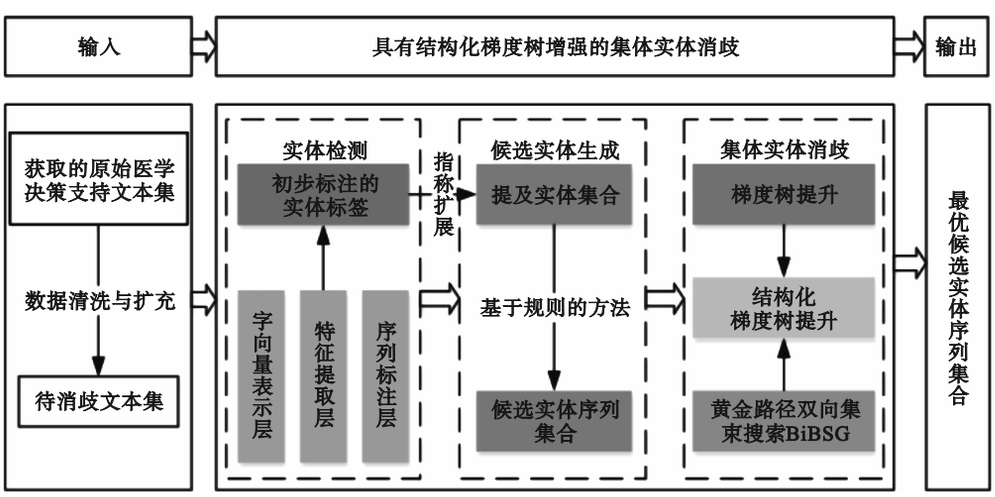


图1　医学决策支持实体消歧总体方案设计图

Fig. 1　Overall scheme design diagram of entity disambiguation for medical decision support

2　具有结构化梯度树提升的集体实体消歧方法

消歧任务是根据提及实体的相关信息和特征对候选实体进行分类和预测,将最优候选实体作为实体消歧的结果.梯度树提升[10]是分类预测任务的重要方法,在进行消歧任务时,分别对提及实体(mention)和候选实体(candidate)进行处理,将损失函数的负梯度值作为近似残差,将先

前残差作为学习目标,多次迭代直至残差达到较小阈值[11].然而,医学领域实体具有专业性和复杂性,实体间彼此关联且具有重要特征.现有应用梯度树提升的实体消歧方法通常是对单个实体进行操作,没有将全部实体进行集体消歧,缺乏全局性[12].

因此,本文针对医疗决策支持信息的专业性、复杂性以及实体间重要特征等问题,提出CED-SGTB方法,从全局角度完成消歧任务.该方法主要由3个部分组成:在实体检测模块中,多个神经网络模型协作对提及实体进行识别和检测,得到标注好的提及实体;在候选实体生成模块中,通过基于规则的候选实体生成方法从提及实体的指称项中获得候选实体;在利用结构化梯度树提升模型进行消歧时,设计BiBSG对候选实体采样,定义全局特征并与局部特征联合建模,通过迭代拟合回归树输出最优的候选实体序列,完成实体消歧.图2给出了该方法的整体结构.

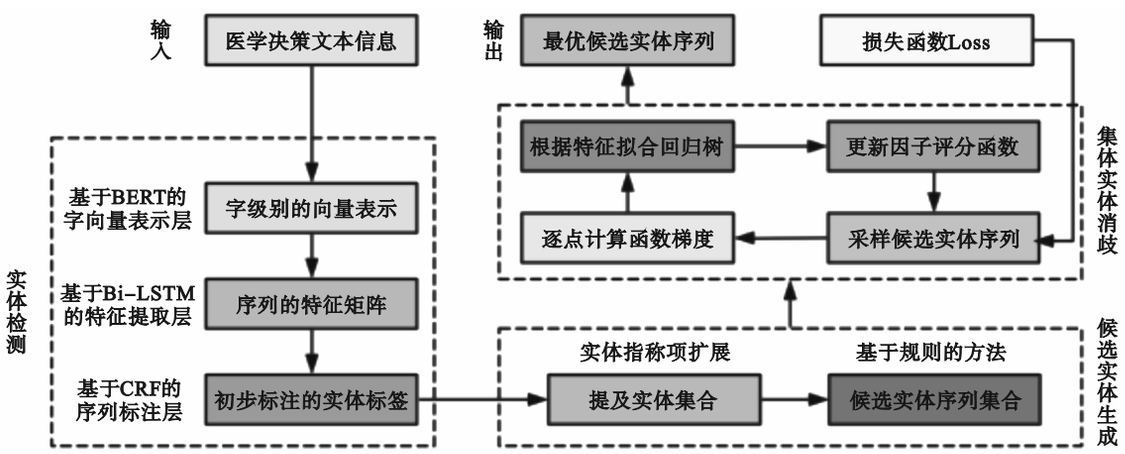


图2 结构化梯度树提升的集体实体消歧方法

Fig. 2 Collective entity disambiguation method with structured gradient tree boosting

2.1 实体检测

实体消歧旨在为文本中的mention匹配最佳语义对应的candidate[13].本文提出多级神经网络协作模型,通过字向量表示层、特征提取层和序列标注层的协同处理,实现医学决策文档的提及检测.

每一个文本x_{ijk}都存在3个特征:字词特征、句子特征及位置特征.对于3种特征的定义如下[14]:

字词特征(token),对每一个x_{ijk}采用通用语义表示模型(BERT)的词汇表确定其对应的字向量;

句子特征(segment),由于本文方法输入的基本单元是句子,所以该特征对整个检测任务的作用不会产生差异,因此设置为0;

位置特征(position),表示该字在句子中的位置,e_{ijk}^{pos}表示句子中第k个字的位置特征.

BERT模型是一种常用的预训练模型,基于BERT模型的字向量表示层示意图如图3所示,通过式(1)计算特征向量.

$$\boldsymbol{E}_k = e_{ijk}^{\mathrm{tok}} + e_{ijk}^{\mathrm{seg}} + e_{ijk}^{\mathrm{pos}}. \tag{1}$$

将图3句子中的每一个x_{ijk}进行以上操作,得到最终向量$\boldsymbol{E}$作为BERT模型的输入,以完成字向量表示.

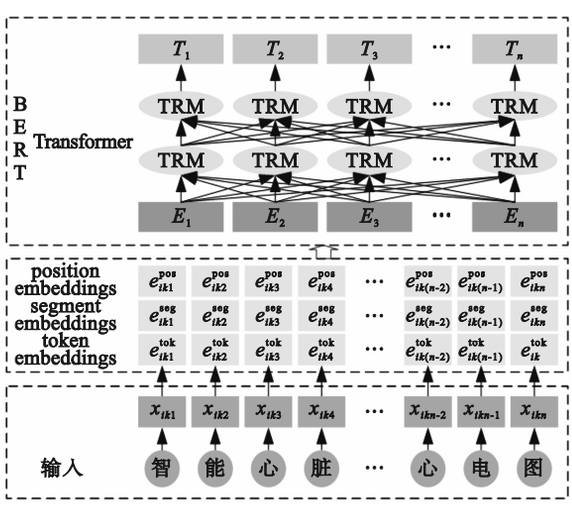


图3 基于BERT模型的字向量表示层

Fig. 3 Word vector representation layer based on BERT model

在特征提取层[15]中,本文选择引入双向长短期记忆网络(Bi-LSTM),通过将上一层的特征输入到当前层,应用长短期记忆网络(LSTM)在相应时间节点进行特征运算.具体流程如图4所示.

如式(2)、式(3)所示,设前向与后向隐藏层输出序列分别为$\mathrm{output}_F=\{F_1,F_2,\cdots,F_n\}$与$\mathrm{output}_B=\{B_1,B_2,\cdots,B_n\}$.

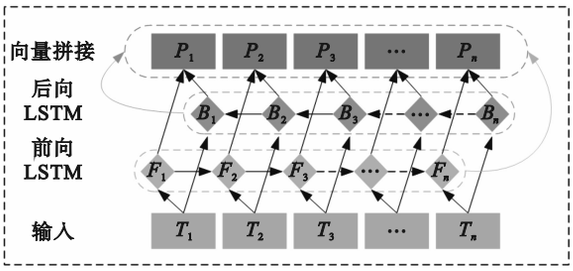


图4　基于Bi-LSTM神经网络的特征提取层

Fig. 4　Feature extraction layer based on Bi-LSTM neural network

$$\text{input}_{F_i}=\begin{cases}T_i, & i=1;\\ T_i\oplus F_{i-1}, & i\geqslant 2.\end{cases} \tag{2}$$

$$\text{input}_{B_j}=\begin{cases}T_j, & j=n;\\ T_j\oplus B_{j+1}, & j\leqslant n-1.\end{cases} \tag{3}$$

其中:⊕表示拼接.每个输入都可以计算出对应的输出向量,如式(4)所示.

$$P_k=F_k\oplus B_k. \tag{4}$$

基于Bi-LSTM的特征提取层的输出是根据输入序列得到的特征矩阵$\boldsymbol{P}$,如式(5)所示.

$$\boldsymbol{P}=\{P_1,P_2,\cdots,P_n\}=\begin{bmatrix}p_{11} & \cdots & p_{1m}\\ \vdots & & \vdots\\ p_{n1} & \cdots & p_{nm}\end{bmatrix}. \tag{5}$$

在基于条件随机场(CRF)的序列标注层[16]中,针对特征提取层的输出矩阵$\boldsymbol{P}$,将输入序列的标记序列定义为$y=\{y_1,y_2,\cdots,y_n\}$,并定义一个输入文档序列与输出结构y之间的联合评分函数$\text{Score}(s_{ij},y)$,如式(6)所示.

$$\text{Score}(s_{ij},y)=\sum_{k=1}^{n}f(s_{ij},y_k,y_{1:(k-1)}). \tag{6}$$

其中:$f(s_{ij},y_k,y_{1:(k-1)})$是因子评分函数[17],每个标签预测值$y_k$都依赖于先前的预测值$y_{1:(k-1)}$,大致流程如图5所示.

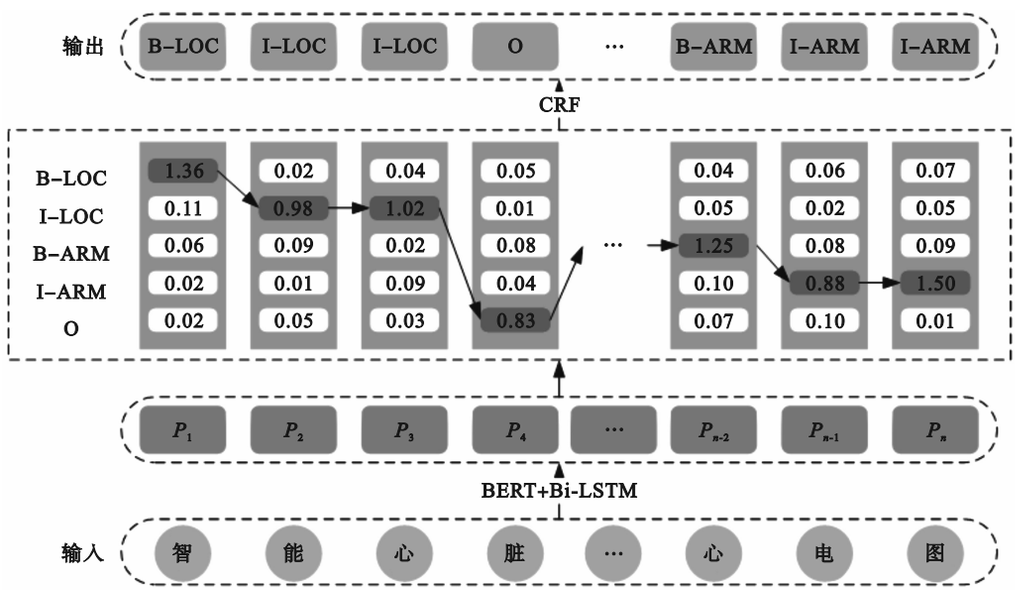


图5　基于CRF的序列标注层

Fig. 5　Sequence annotation layer based on CRF

2.2　候选实体生成

实体检测阶段首先识别并标注输入文档中的提及实体mention,但要实现准确的实体链接,需要从知识库中提取该mention可能对应的所有候选指称项candidate,为后续的实体消歧任务提供支持[18].

较短的实体指称项往往存在更高的歧义性[19],因此需要通过指称扩展对文本中的别名、缩写等表述进行规范化处理.本研究采用上下文信息与知识库相结合的方式实现指称扩展,具体流程详见图6.

首先进行指代消解,获取实体指代链;若指称在链上,则加入最长实体,否则判断是否为缩略词,是则加入全称,否则加入上下文中最长的两个相关实体或直接加入指称[20],在知识库中检索修正实体,选取高置信度结果,经预处理后得到标准实体集合.本文采用的基于规则的候选生成方法如表1所示.

采用分级规则从知识库检索候选实体,逐级填充候选集合.若5级规则均未命中,则返回空集.该方法可为每个mention生成对应候选集,从而提升覆盖率和消歧可靠性.

2.3　实体集体消歧

得到mention集合和candidate集合后,开始进行消歧.对实体检测部分的联合评分函数进行softmax归一化处理,如式(7)、式(8)所示.

$$p(y|s_{ij})=\frac{\exp(\text{Score}(s_{ij},y))}{Z(s_{ij})},\tag{7}$$

$$Z(s_{ij})=\sum_{y'\in \text{Dict}(s_{ij})}\exp\left(\sum_{k=1}^{n}f(s_{ij},y'_k,y'_{1:(k-1)})\right).\tag{8}$$

其中：$p(y|s_{ij})$是可能输出结构的分布；$Z(s_{ij})$是一个全局归一化项；$\text{Dict}(s_{ij})$是根据词典确定的所有可能候选实体序列的集合.

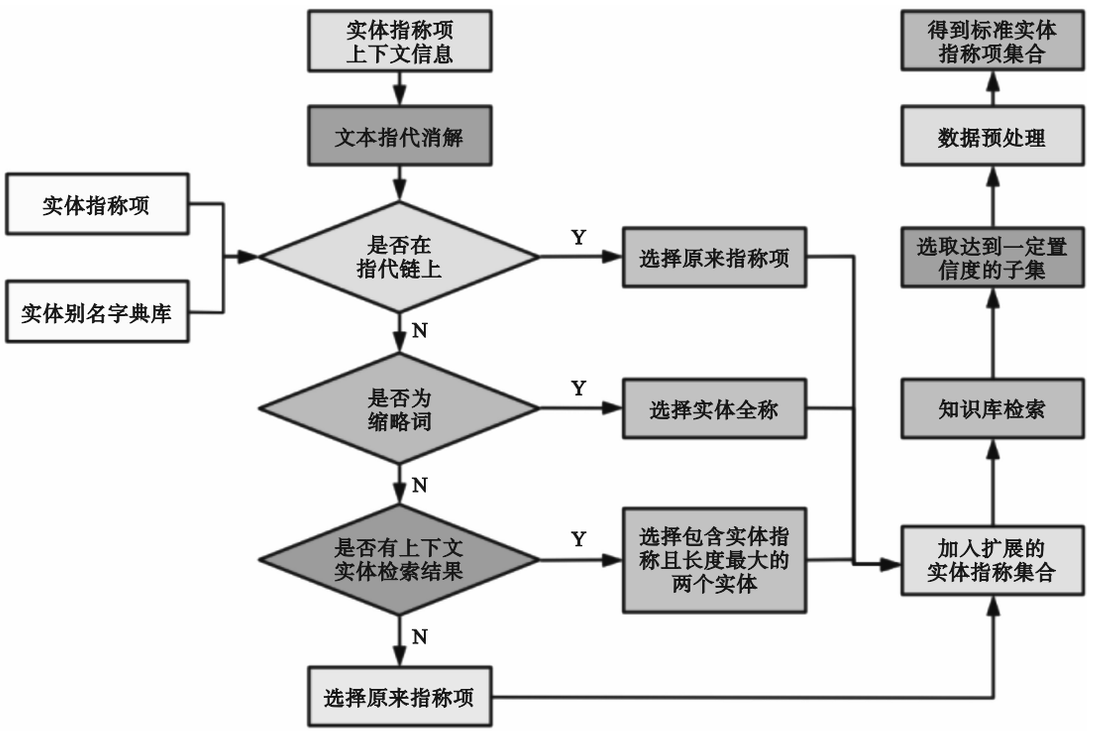


图6 结合上下文与知识库的指称项扩展方法流程

Fig. 6 Process of referential item extension method combining context and knowledge base

表1 基于规则的候选实体生成方法

Table 1 Rule-based candidate entity generation method

级别	规则详述
1	完全与mention名称/昵称/别名相匹配
2	完全与mention名称/昵称/别名的拼音相匹配
3	基于编辑距离，部分与mention名称/昵称/别名相匹配
4	整体包含mention名称/昵称/别名
5	包含mention名称/昵称/别名的一部分文字

消歧任务的最终目标就是找到最优的输出序列[21]，如式(9)所示.

$$y^{*}=\underset{y\in \text{Dict}(s_{ij})}{\text{argmax}}\ p(y|s_{ij})=\underset{y\in \text{Dict}(s_{ij})}{\text{argmax}}\sum_{k=1}^{n}f(s_{ij},y_k,y_{1:(k-1)}).\tag{9}$$

其中：argmax是求函数参数的函数，得到目标函数取最大值时的函数集合.

本文改进传统梯度树提升模型[22]，提出结构化梯度树提升，通过结合实体上下文局部特征和历史决策全局特征，使用回归树进行集体消歧. 选用输入序列与输出结构间条件概率的负对数似然函数[23]作为损失函数(Loss)，通过多次迭代使其最小化，如式(10)所示.

$$\text{Loss}=L(y^{*},\text{Score}(s_{ij},y))=-\lg p(y^{*}|s_{ij})=\lg Z(s_{ij})-\text{Score}(s_{ij},y).\tag{10}$$

其中：y^{*}是最优输出结构，即黄金输出结构. 模型每一次迭代都要经历4个步骤.

1) 采样候选实体序列. 在候选实体生成阶段，计算mention的前20个候选实体时会产生较高的时间复杂度，影响训练效率. 因此，采用集束搜索算法[24]，通过在每次迭代中保留多个条件概率较大的输出序列，并将先前结果作为下一个时间步的输入，从而提高搜索精度，降低计算开销.

通过集束搜索在解码训练序列时对集束中的黄金输出结构进行追踪，并在其脱离集束或到达最后一步时进行梯度更新. 但每个训练阶段只更新一次，逐点计算梯度较为烦琐. 为此，本文提出黄金路径集束搜索算法，该算法可以避免黄金输出结构脱离集束，确保每个时间步长都能获得有效的梯度更新，从而提高模型训练和使用效率.

搜索采样时，考虑到在时间步长t所做的决策情况，将联合概率$p(y,s_{ij})$在t处进行分解，如式(11)所示.

$$p(y|s_{ij})=p(y_{1:(t-1)}|s_{ij})\times p(y_t|y_{1:(t-1)},s_{ij})\times p(y_{(t+1):T}|y_t,y_{1:(t-1)},s_{ij}).\tag{11}$$

传统的集束搜索只是从左至右的单向推理,在时间步长 t 内只能考虑当前的集束序列 $y_{1:(t-1)}$,后续序列是通过假设其可均等概率的方式来近似的.针对该问题,本文采用双向集束搜索算法,合并多个集束从而充分计算和考虑未来信息,如式(12)所示.

$$p(y_{(t+1):T}|y_t,y_{1:(t-1)},s_{ij})=p(y_{(t+1):T}|y_t,s_{ij})\propto p(y_t|y_{(t+1):T},s_{ij})\cdot p(y_{(t+1):T}|s_{ij}). \tag{12}$$

其中:$\propto$ 表示前者与后者呈正比例关系.由此可得新的条件概率表达式,如式(13)所示.通过联合评分函数计算正向和反向集束部分序列的概率分布.

$$p(y|s_{ij})\propto p(y_{1:(t-1)}|s_{ij})\times p(y_t|y_{1:(t-1)},s_{ij})\times p(y_t|y_{(t+1):T},s_{ij})\times p(y_{(t+1):T}|s_{ij}). \tag{13}$$

因此,本文在采样候选序列过程中,对集束搜索进行改进,提出黄金路径双向集束搜索算法,充分利用过去和未来的信息减小整体模型的方差,实现更优的局部搜索.

2) 逐点计算函数梯度.结构化梯度树提升模型在函数空间中执行梯度下降,从而迭代因子评分函数 $f(\cdot)$.在迭代前将 $f(\cdot)$ 初始化为 0,在第 m 次迭代中,$f(\cdot)$ 的更新结果如式(14)所示.

$$f_m(s_{ij},y_t,y_{1:(t-1)})=f_{m-1}(s_{ij},y_t,y_{1:(t-1)})-\eta_m g_m(s_{ij},y_t,y_{1:(t-1)}). \tag{14}$$

其中:g_m 为函数梯度;η_m 为学习速率.若要更新 $f(\cdot)$,需要逐点计算采样候选序列的函数梯度,如式(15)所示.

$$g_m(s_{ij},y_t,y_{1:(t-1)})=\frac{\partial L(y^*,\mathrm{Score}(s_{ij},y))}{\partial f(s_{ij},y_t,y_{1:(t-1)})}=p(y_{1:t}|s_{ij})-\mathrm{ind}[y_{1:t}=y^*_{1:t}]. \tag{15}$$

式中:ind[·]是指标函数 indicator,若预测序列为黄金序列[25]则返回 1,否则返回 0.

由于函数梯度的计算过程在不同文本之间是相互独立的,因此在每次训练迭代时,将训练文本集随机划分为不同的分区,并行计算文本的逐点函数梯度值,以加快训练效率.

3) 拟合回归树.因子评分函数通常被假定为 $f(s_{ij},y_t,y_{1:(t-1)})=\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\varphi}(s_{ij},y_t,y_{1:(t-1)})$ 的形式,其中 $\boldsymbol{\theta}$ 是模型参数.$\boldsymbol{\varphi}(s_{ij},y_t,y_{1:(t-1)})$ 是特征函数,可以分解为局部特征函数 $\boldsymbol{\varphi}_{\mathrm{local}}(s_{ij},y_t)$ 和全局特征函数 $\boldsymbol{\varphi}_{\mathrm{global}}(y_t,y_{1:(t-1)})$ 两部分[26].

对于 $\boldsymbol{\varphi}_{\mathrm{local}}(s_{ij},y_t)$,本文分别采用以下 3 种局部特征指标:

①实体流行度特征 $\mathrm{pop}_{\mathrm{ent}}$:目标实体被链接过的次数越多,则流行度越高.流行度是候选序列中 candidate 被链接次数 c_i 占总链接次数的比例,计算方法如式(16)所示.

$$\mathrm{pop}_{\mathrm{ent}}=\frac{c_i}{\sum_{i=1}^{n}c_i}. \tag{16}$$

②字符串相似度特征 $\mathrm{sim}_{\mathrm{str}}$:采用基于编辑距离的方法,用输入文档和候选序列中的 mention 与 candidate 的编辑距离与 mention 的长度作比,如式(17)所示.

$$\mathrm{sim}_{\mathrm{str}}=\frac{d[\mathrm{mention}_i,\mathrm{candidate}_{ij}]}{|\mathrm{mention}_i|}. \tag{17}$$

③类型相似度特征 $\mathrm{sim}_{\mathrm{type}}$:通过实体检测对实体类型进行标注,将输入文档中 mention 类型集合与候选序列中 candidate 类型集合进行余弦相似度计算,如式(18)所示.

$$\mathrm{sim}_{\mathrm{type}}=\cos(m_i,c_{ij})=\frac{m_i\cdot c_{ij}}{\|m_i\|\cdot\|c_{ij}\|}. \tag{18}$$

对于全局特征,本文考虑当前候选实体序列与之前实体决策历史记录的关系,定义实体-实体特征向量 $\boldsymbol{\varphi}_{\mathrm{ent}}(y_t,y_{t'})$ 为实体的共现次数以及候选实体序列之间的余弦相似度得分.为提高余弦相似度得分的计算速率,通过将候选实体嵌入归一化,再利用点积运算得到相似性特征.在一个时间步长为 t 的训练序列中,使用历史决策 $y_{1:t-1}$ 来对 y_t 的一致性进行量化,得到长度固定的全局特征函数 $\boldsymbol{\varphi}_{\mathrm{global}}(y_t,y_{1:(t-1)})$,如式(19)所示.

$$\boldsymbol{\varphi}_{\mathrm{global}}(y_t,y_{1:(t-1)})=\sum_{t'}^{t-1}\frac{\boldsymbol{\varphi}_{\mathrm{ent}}(y_t,y_{t'})}{t-1}\oplus\max_{t'}^{t-1}\boldsymbol{\varphi}_{\mathrm{ent}}(y_t,y_{t'}). \tag{19}$$

通过回归树模型 $h_m(\cdot)$ 近似负函数梯度 $-g_m(\cdot)$,该负函数梯度可以通过将训练数据的特征函数 $\boldsymbol{\varphi}(s_{ij}^{(k)},y_t^{(k)},y_{1:(t-1)}^{(k)})$ 拟合到负函数梯度的对应点 $-g_m(s_{ij}^{(k)},y_t^{(k)},y_{1:(t-1)}^{(k)})$ 得到.

4) 更新因子评分函数.在完成以上步骤后,使用式(20)更新第 m 次迭代的因子评分函数 $f(\cdot)$.

$$f(s_{ij},y_t,y_{1:(t-1)})=\sum_{m=1}^{M}\eta_m h_m(s_{ij},y_t,y_{1:(t-1)}). \tag{20}$$

其中:$h_m(s_{ij}^{(k)},y_t^{(k)},y_{1:(t-1)}^{(k)})$ 是一个基函数;η_m 为学习速率.

根据因子评分函数对候选实体序列进行采样,重复上述步骤,直到损失函数达到最小值.此时得到的候选序列即为最优输出序列 y^*.

3　实验验证与结果分析

3.1　实体消歧验证实验

1) 实验数据集.本文实验数据主要采用与医学决策支持相关的文本数据,通过爬取医学领域

的书籍和相关网络报道以及提取的医学百科数据,经过基本清洗和扩充后,得到本文的实验数据集,按照7:2:1的比例将数据集分为训练集、开发集和测试集[27].在训练期间,每10个epoch检查一次开发集的性能,以执行早期停止策略[28].数据集信息如表2所示.

表2 数据集信息

Table 2 Dataset information

数据集划分	具体数量
句子总数	12 607
训练集	8 825
开发集	2 521
测试集	1 261

2)模型评价指标.为了更好地验证本文模型的性能,选取精确度(ACC)[29]和调和平均值($F1$)[30]两个评价指标.这两个指标的计算均需要依托混淆矩阵中的值,混淆矩阵如表3所示.

表3 混淆矩阵

Table 3 Confusion matrix

预测结果	正例	负例
正例	TP	FP
负例	FN	TN

ACC的计算方法如式(21)所示,即预测正确的数量占总体数量的比值.

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}. \tag{21}$$

计算综合评价指标$F1$需要先计算准确率(Precision)和召回率(Recall),如式(22)、式(23)所示.Precision是指正确预测为正例的数量占所有预测为正例的总数量的比值;Recall是指正确预测为正例的数量占真实正例数量的比值.

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{22}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{23}$$

$F1$的计算方法如式(24)所示.

$$F1=\frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}. \tag{24}$$

3.2 仿真实验与性能分析

3.2.1 模型参数影响实验

1)回归树最大深度对模型影响实验.模型通过拟合回归树迭代更新参数,当回归树最大深度不同时,会产生不同的实验结果.将回归树最大深度设置为不同数值进行实验,结果如图7所示.

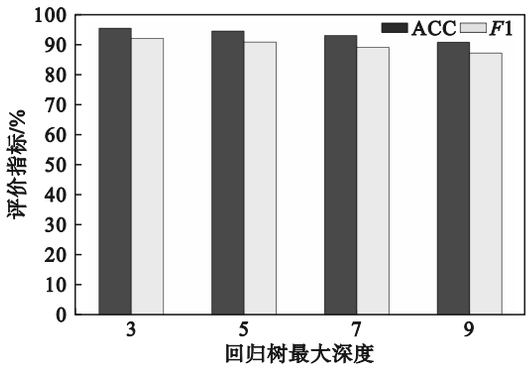


图7 回归树最大深度对模型性能的影响

Fig. 7 Effect of maximum depth of regression tree on model performance

由实验结果可知,回归树的最大深度会对实验结果造成影响.当将其设置为3时,可以取得更好的性能,所以在后续实验中将最大深度设置为3.

2)集束尺寸对模型的影响实验.集束搜索在每个时间步长保留多个输出,改进贪心搜索仅保留最大概率输出的方式,提高了搜索精度.保证其他参数不变,通过调整集束尺寸进行实验,结果如图8所示.

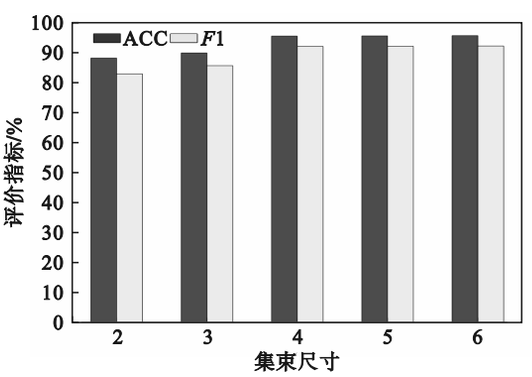


图8 集束尺寸对模型性能的影响

Fig. 8 Effect of beam size on model performance

由实验结果可知,当集束尺寸过小时,模型性能较差;当集束尺寸过大时,虽然性能有所提升,但会增加较大的计算代价,影响模型运行效率.当设置为4时,可得到较好的性能且时间复杂度较低.因此,在后续实验中将集束尺寸设置为4.

3.2.2 模型稳定性实验

为了验证模型的稳定性,分别从数据集中抽取5 000,6 000,7 000,8 000和9 000条数据作为训练集,并且统一设置2 000条作为开发集、1 000条作为测试集进行实验,对ACC和$F1$两个指标进行评估,结果如表4所示.

表4　不同数据量下的消歧性能变化

Table 4　Disambiguation performance changes under different data volumes

数据量/条	ACC/%	*F*1/%
5 000	85. 43	81. 84
6 000	89. 01	85. 06
7 000	92. 66	88. 92
8 000	94. 85	90. 67
9 000	95. 30	91. 88

由结果可知,当数据集较小时,训练集与开发集、测试集的比例也较小,很难获得较好的学习效果.而随着训练数据的增多,ACC 和 *F*1 值都逐渐增大.这说明本文模型在数据量较大时能得到更好的效果,具有一定的稳定性和可靠性.

3. 2. 3　模型对比实验

1) 不同实体检测方法对模型性能影响对比实验.为验证本文采用的 BERT+Bi-LSTM+CRF 方法对消歧任务的提升效果,分别将其与其他方法得到的标注结果作为后续步骤的输入,对性能进行评估,实验结果如表 5 所示.

表5　不同实体检测方法对消歧性能的影响

Table 5　Effect of different entity detection methods on disambiguation performance　%

实体检测方法	ACC	*F*1
CRF	76. 54	73. 17
Bi-LSTM+CRF	82. 63	77. 91
CNN+Bi-LSTM+CRF	89. 39	85. 22
BERT+Bi-LSTM+CRF	**95. 27**	**91. 85**

实验结果表明,本文提出的基于多神经网络协作的实体检测方法优于传统方法.与 CRF 方法相比,本文通过 Bi-LSTM 更准确地提取文本特征;与 Bi-LSTM+CRF 方法相比,采用 BERT 更好地进行文本向量化;与 CNN+Bi-LSTM+CRF 方法相比,本文充分考虑了句子和位置特征,并使用多层 Transformer 训练字向量,表现更佳.因此,本文方法在实体检测阶段具有显著优势,有助于提升后续消歧效果.

2) 不同实体消歧模型的性能对比实验.本文在实体检测和候选实体生成后提出基于结构化梯度树提升模型的集体消歧方法.为了验证模型的优异性,将本文模型与其他消歧模型进行对比,选取以下 3 种作为实验的对比模型,对不同训练轮次(epoch)下的 *F*1 分数进行评估,得到的实验结果如图 9 所示.

Neu-PL:使用 LSTM 和注意力机制消除歧义,有效测量提及实体上下文和目标实体之间的语义匹配.

RRWEL:建立局部上下文特征模型,使用随机游走层堆叠,全局考虑不同消歧决策之间的语义依赖.

HRFAENE:同时考虑特征网络结构信息和文档属性特征.

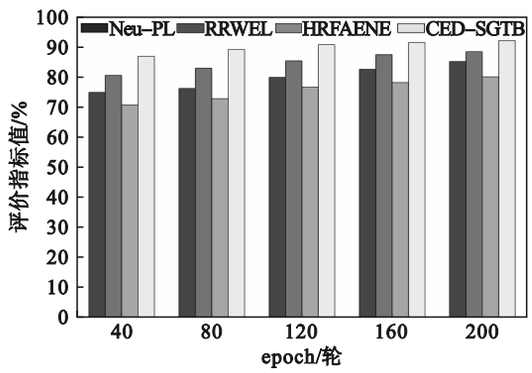


图9　不同消歧模型在不同 epoch 下的 *F*1 分数对比

Fig. 9　Comparison of *F*1 scores for different disambiguation models at different epochs

实验结果表明,CED-SGTB 模型在不同 epoch 下表现优异,其 *F*1 分数领先其他模型.这是因为本文模型能够通过全局考虑文档结构信息,结合黄金路径双向集束搜索算法采样候选实体序列,并联合建模局部特征与全局信息,迭代更新参数,进而提升消歧效果.

3. 2. 4　模型消融实验

本文模型主要通过改进黄金路径双向集束搜索算法来采样候选实体;另外在传统梯度树提升的基础上,利用实体上下文的局部特征与消歧历史决策的全局特征联合建模,实现全局消歧.为探究各部分对模型性能的影响,进行消融实验,相关简化模型如下:

GTB,传统梯度树提升模型,未对局部特征与全局特征进行联合建模;

SGTB-BS,仅采用集束搜索算法对候选实体序列进行采样;

SGTB-BSG,采用黄金路径集束搜索算法对候选实体序列进行采样;

SGTB-BiBSG,完整模型,采用黄金路径双向集束搜索对候选实体序列进行采样.

对实验结果的两个评价指标进行评估分析,实验结果如表 6 所示.

由实验结果可知,本文完整模型在两个评价指标上均优于简化模型.对比实验 1 和实验 4,由

于本文模型能够将全局信息与局部信息联合建模,实现全局优化,所以在ACC和$F1$值上领先GTB模型;对比实验2,3,4,在传统集束搜索基础上改进为黄金路径集束搜索可以在一定程度上提高消歧精度,分别在ACC和$F1$上取得提升;黄金路径双向集束搜索算法能在实验3得到的性能效果基础上得到更优的精确度和$F1$.综合4组实验来看,本文的改进算法与结构化梯度树提升均能为消歧任务提供支持和促进.

表6 不同简化模型的消歧性能结果对比

Table 6 Comparison of disambiguation performance results for different simplified models %

序号	模型变体	ACC	$F1$
1	GTB	91.40	89.07
2	SGTB-BS	92.08	89.73
3	SGTB-BSG	93.19	90.52
4	SGTB-BiBSG	**95.27**	**91.85**

4 结 语

本文围绕医学知识图谱构建中的知识融合技术,重点研究了实体消歧方法.针对医学数据的专业性和复杂性,提出了基于结构化梯度树提升的集体实体消歧方法(CED-SGTB).通过系列实验验证:

1)确定回归树深度3和集束尺寸4为最优参数;

2)模型在8 000,9 000条训练数据量时趋于稳定;

3)对比实验表明该模型在特征提取和全局建模方面优于其他方法;

4)消融实验证实,改进的BiBSG算法和双向集束搜索显著提升了消歧精度.

研究成果为医学知识图谱构建提供了有效的技术支持,有助于提升医疗决策的准确性和效率.

参考文献:

[1] Körschens M, Bucher S F, Bodesheim P, et al. Determining the community composition of herbaceous species from images using convolutional neural networks [J]. *Ecological Informatics*, 2024, 80: 102516.

[2] Qian Y, Pan L. Leveraging multimodal features for knowledge graph entity alignment based on dynamic self-attention networks [J]. *Expert Systems with Applications*, 2023, 228: 120363.

[3] Ran Y Y, Xu X B, Luo M Z, et al. Scene classification method based on multi-scale convolutional neural network with long short-term memory and whale optimization algorithm[J]. *Remote Sensing*, 2024, 16(1): 174.

[4] Liu X, Zhang L, Zheng Q S, et al. Construction of an event knowledge graph based on a dynamic resource scheduling optimization algorithm and semantic graph convolutional neural networks [J]. *Electronics*, 2024, 13(1): 11.

[5] Nafa Y, Chen Q, Hou B Y, et al. Adaptive deep learning for entity disambiguation via knowledge-based risk analysis [J]. *Expert Systems with Applications*, 2024, 238: 122342.

[6] Basile A, Crupi R, Grasso M, et al. Disambiguation of company names via deep recurrent networks [J]. *Expert Systems with Applications*, 2024, 238: 122035.

[7] Wu W M, Hu J T, Zhang Z J, et al. Deterministic learning-based neural identification and knowledge fusion [J]. *Neural Networks*, 2023, 166: 688-702.

[8] Wei F F, Mei K Z. Frequency inception based graph neural network for relation prediction in knowledge graphs [J]. *Knowledge-Based Systems*, 2023, 278: 110908.

[9] 杨光, 刘秉权, 刘铭. 基于图方法的命名实体消歧[J]. 智能计算机与应用, 2015, 5(5): 52-55.
(Yang Guang, Liu Bing-quan, Liu Ming. Named entity disambiguation based on graph method [J]. *Intelligent Computer and Applications*, 2015, 5(5): 52-55.)

[10] Yang L H, Chen J R, Wang Z H, et al. Subgraph-aware virtual node matching graph attention network for entity alignment [J]. *Expert Systems with Applications*, 2023, 231: 120694.

[11] Guo S D, Liu X J, Zhang H Y, et al. Causal knowledge fusion for 3D cross-modality cardiac image segmentation [J]. *Information Fusion*, 2023, 99: 101864.

[12] Wang S, Zhang Y, Hu Y L, et al. Knowledge fusion enhanced graph neural network for traffic flow prediction [J]. *Physica A: Statistical Mechanics and Its Applications*, 2023, 623: 128842.

[13] Taha K. Semi-supervised and un-supervised clustering: a review and experimental evaluation [J]. *Information Systems*, 2023, 114: 102178.

[14] Matsui A, Murakami M. Deferred acceptance algorithm with retrade[J]. *Mathematical Social Sciences*, 2022, 120: 50-65.

[15] 张吉祥, 张祥森, 武长旭, 等. 知识图谱构建技术综述 [J]. 计算机工程, 2022, 48(3): 23-37.
(Zhang Ji-xiang, Zhang Xiang-sen, Wu Chang-xu, et al. Review of knowledge graph construction techniques [J]. *Computer Engineering*, 2022, 48(3): 23-37.)

[16] Wang C X, Huang Z H, Wan Y, et al. FuAlign: cross-lingual entity alignment via multi-view representation learning of fused knowledge graphs [J]. *Information Fusion*, 2023, 89: 41-52.

[17] Bouarroudj W, Boufaida Z, Bellatreche L. Named entity disambiguation in short texts over knowledge graphs [J]. *Knowledge and Information Systems*, 2022, 64(2): 325-351.

[18] Zhang Z R, Yang Y Y, Chen B H. Relation-aware heterogeneous graph neural network for entity alignment [J]. *Neurocomputing*, 2024, 592: 127797.

[19] Yang S Y, Zhang P L, Che C, et al. B-LBConA: a medical entity disambiguation model based on Bio-LinkBERT and context-aware mechanism [J]. *BMC Bioinformatics*, 2023, 24(1): 97.

[20] Zhu B B, Bao T, Wang K R, et al. A semi-supervised neighborhood matching model for global entity alignment [J]. *Neural Computing and Applications*, 2023, 35(15): 10779-10799.

(下转第29页)