计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(1):280−288 [doi: 10.15888/j.cnki.csa.009370]

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

基于 RoBERTa 和 T5 的两阶段医学术语标准化①

周　景[1], 崔灿灿[1], 王梦迪[2], 王泽敏[1]

[1](华北电力大学 控制与计算机工程学院, 北京 102206)
[2](北京中科睿见科技有限公司, 北京 100080)
通信作者: 周　景, E-mail: zhoujing108@ncepu.edu.cn

摘　要: 医学术语标准化作为消除实体歧义性的重要手段, 被广泛应用于知识图谱的构建过程之中. 针对医学领域涉及大量的专业术语和复杂的表述方式, 传统匹配模型往往难以达到较高的准确率的问题, 提出语义召回加精准排序的两阶段模型来提升医学术语标准化效果. 首先在语义召回阶段基于改进的有监督对比学习和 RoBERTa-wwm 提出语义表征模型 CL-BERT, 通过 CL-BERT 生成实体的语义表征向量, 根据向量之间的余弦相似度进行召回并得到标准词候选集, 其次在精准排序阶段使用 T5 结合 prompt tuning 构建语义精准匹配模型, 并将 FGM 对抗训练应用到模型训练中, 然后使用精准匹配模型对原词和标准词候选集分别进行精准排序得到最终标准词. 采用 ccks2019 公开数据集进行实验, *F*1 值达到了 0.920 6, 实验结果表明所提出的两阶段模型具有较高的性能, 为实现医学术语标准化提供了新思路.

关键词: 医学术语标准化; RoBERTa-wwm; 对比学习; T5; prompt tuning; 知识图谱

引用格式: 周景,崔灿灿,王梦迪,王泽敏.基于 RoBERTa 和 T5 的两阶段医学术语标准化.计算机系统应用,2024,33(1):280–288. http://www.c-s-a.org.cn/1003-3254/9370.html

Two-stage Medical Terminology Standardization Based on RoBERTa and T5

ZHOU Jing[1], CUI Can-Can[1], WANG Meng-Di[2], WANG Ze-Min[1]

[1](School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China)
[2](Beijing Smart Insight Technology Co. Ltd., Beijing 100080, China)

Abstract: Medical terminology standardization, as an important means to eliminate entity ambiguity, is widely used in the process of building knowledge graphs. Aiming at the problem that the medical field involves a large number of professional terminology and complex expressions, and the traditional matching models are often difficult to achieve a high accuracy rate, a two-stage model of semantic recall and precise sorting is proposed to improve the standardization effect of medical terminology. First, in the semantic recall stage, a semantic representation model CL-BERT is proposed based on the improved supervised contrastive learning and RoBERTa-wwm. The semantic representation vector of an entity is generated through CL-BERT, and recall is carried out according to the cosine similarity between the vectors, so as to obtain the standard word candidate set. Secondly, in the precise sorting stage, T5, combined with prompt tuning, is used to build a precise semantic matching model, and FGM confrontation training is applied to the model training; next, the precise matching model is used to precisely sort the original word and standard word candidate sets, so as to obtain the final standard words. The ccks2019 public data set is used for experiments, achieving an *F*1 value of 0.920 6. The experimental results show that the proposed two-stage model showcases high performance, and provides a new idea for medical terminology standardization.

Key words: medical terminology standardization; RoBERTa-wwm; contrastive learning; T5; prompt tuning; knowledge graph

① 收稿时间: 2023-05-18; 修改时间: 2023-06-26, 2023-08-17; 采用时间: 2023-08-31; csa 在线出版时间: 2023-11-24
CNKI 网络首发时间: 2023-11-28

1 引言

在信息爆炸式增长的今天, 患者很难在短时间内从互联网大量的冗余信息中找到自己想要的医疗与病理信息, 因此搭建一个能够满足人们日常医疗所需的医疗问答系统成为当下研究的热点. 医疗知识图谱作为医疗问答系统的重要的知识源, 是构建准确、全面医疗问答系统的重要基础. 在医疗知识图谱的构建中, 由于不同的医生可能对同一种疾病所使用的表达方式不同, 导致医学术语具有很高的歧义性, 如何消除实体歧义性成为当下构建医药知识图谱的一大难点. 使用医学术语标准化技术可以将不同的医生对同一种疾病的表述进行规范化, 从而消除歧义, 提高医药知识图谱的准确性和全面性. 本质上, 医学术语标准化属于实体链接任务. 本文以《ICD9-2017 协和临床版》标准术语库为例, 在该标准术语库中找到一个或多个标准术语与电子病历中提到的医学术语进行对齐. 比如将电子病历中所描述“右股骨病灶活检术”映射为标准化术语库中的“股骨活组织检查”.

早期的医学术语标准化任务多采用基于规则和机器学习实现. Ghiasvand 等人[1]提出基于编辑距离特征来生成候选集的方法, 提升了术语标准化的性能. Kang 等人[2]使用包含 5 种规则的自然语言处理模块, 有效提升了生物医学文本中的疾病规范化效果. Leaman 等人[3]尝试使用机器学习模型 DNorm 处理医学术语标准化任务, 证明了机器学习解决这一类问题的可行性. 但基于机器学习的方法只依赖于两个实体之间的字面差异, 未涉及实体之间的深层语义表征.

近年来, 随着深度学习的发展, 基于深度学习的自然语言处理技术在医学术语标准化任务上表现不断突破. Li 等人[4]将生物医学实体标准化的任务转换为交互式语义匹配排序问题, 构建卷积神经网络 (CNN) 对生物医学实体对进行打分排序, 该排序方法优于传统的基于规则的方法. Luo 等人[5]提出多任务框架, 使用多视角 CNN 来进行特征提取, 并引入权重共享层来处理医学术语标准化任务. Zhang 等人[6]通过无监督学习的方法对标准化任务执行时间进行优化, 相对于传统有监督学习的标准化任务所耗费的时间更短. Tutubalina 等人[7]使用双向 LSTM+GRU 结合进行特征提取, 之后再与 UMLS 系统中的标准化术语链接评分, 相较于基于卷积神经网络来处理医学术语标准化任务取得了更好的结果. 赵兰枝等人[8]使用基于卷积神经网络的模型来研究实体标准化, 标准实体由向量空间模型处理成为标准向量, 输入文本中的通俗实体经由卷积神经网络提取其中的语义特征并转化成为特征向量, 采用 1 个卷积层和 2 个全连接层的浅层网络结构降低模型的复杂程度.

随着 BERT 语言模型的兴起, 采用预训练模型 BERT 及其变体的方法逐渐成为主流. Ji 等人[9]通过微调预训练 BERT/BioBERT/ClinicalBERT 模型提出了实体归一化架构, 首先通过 BM25 模型从标准实体库中进行相似度召回, 召回 top10 作为候选实体, 然后使用 BERT 预训练语言模型对实体文本对编码并进行 0, 1 分类. 分类结果作为标准化结果输出. 实验结果表明对预训练的语言表示模型进行微调, 能够有效地提升生物医学命名实体规范化的水平. Kalyan 等人[10]提出基于 RoBERTa 和联合学习的方法实现医学概念标准化, 在 3 个标准数据集上优于现有方法, 准确率提高 2.31%. Chen 等人[11]提出了一种轻量级的生物医学实体链接神经网络, 只需要使用 BERT 模型的一小部分参数和更小的计算资源, 该方法使用带有注意力机制的简单对齐层来捕捉输入实体和候选实体名称之间的变化. 胡宇等人[12]提出了一种深度学习和知识库相结合的实体链接方法, 通过深度挖掘自然语言文本的隐藏特征, 及其与知识库概念图间结构的相似性, 将实体对齐和实体识别任务统一处理. Li 等人[13]引入知识库和数据增强提升训练样本多样性, 然后使用 BM25 生成候选概念, 最后通过设计的 stacking-BERT 模型捕获语义信息并使用堆叠机制选择最优映射对, 所提出的 stacking-BERT 模型表现优于单一的 BERT 模型和其他传统的深度学习模型. 闫璟辉等人[14]将医学术语标准化任务看作是翻译任务, 第 1 阶段使用生成式模型生成候选实体, 第 2 阶段使用 BERT 预训练模型, 对候选实体进行语义相似度排序, 获得最终的医学术语标准化结果, 该方案在医学术语标准化任务中取得了较好的效果. 韩振桥等人[15]在医学术语标准化任务中使用多策略召回与 RoBERTa 语义打分模型结合的两阶段方案. 第 1 阶段基于传统的 Jaccard、TF-IDF 等多种统计学方法进行相似度召回. 第 2 阶段使用 RoBERTa-wwm-ext 预训练语言模型对待匹配实体和候选实体进行句子对文本分类, 根据分类结果判断两个医学术语是否可以对齐, 在实际应用中取得了不错的效果.

综上, 基于规则的方法需要针对不同场景设定大量规则, 耗时费力且可移植性差. 基于机器学习的方法由于缺乏语义和上下文信息的限制, 难以胜任更复杂的医学术语标准化任务. 基于深度学习的方法能够自动提取特征, 避免了消耗过多的人力设计规则和特征, 同时在文本建模方面表现出强大的表征能力, 能够学习到词语的上下文信息. 随着 BERT 预训练语言模型的兴起, 因其在上下文中能够获得更丰富的语义特征, 在自然语言处理任务中表现出色, 因此使用预训练模型实现医学术语标准化任务已经成为主流.

基于预训练模型的医学术语标准化任务两阶段方法中, 第 1 阶段基于机器学习的方法只依赖于每个字在文本中出现的频率统计, 未涉及实体的语义特征. 对于使用生成式模型生成候选标准词的方案, 生成的候选词可能并不是标准术语库中的词. 在第 2 阶段基于深度学习的方案多使用 BERT、RoBERTa 等预训练语言模型将待匹配实体和候选实体集分别进行文本对语义匹配, 这种传统 fine tuning 的方案下游任务与预训练任务不能够保持一致, 导致模型预训练阶段学到的丰富的语言知识不能充分地应用到下游任务中. 针对上述出现的问题, 本文提出语义召回加精准排序的两阶段方案进一步提升模型效果.

2 模型介绍

对于本文的医学术语标准化任务定义为: 设标准术语数量为 k, 其中标准术语集为 $G=\{g_1, g_2, \cdots, g_k\}$, 术语原词为 s, 在标准术语集 G 中找到原词 s 所对应的标准词.

本文使用语义召回加精准排序的两阶段方案. 在语义召回阶段: 通过 CL-BERT 将术语原词和标准术语库中全部医学术语转化为语义向量, 计算原词向量和所有标准术语向量的余弦相似度, 召回余弦相似度最高的前 10 个标准术语向量, 并将其与标准术语库中的词一一对应, 作为标准词候选集. 在精准排序阶段: 将候选标准词与原词分别组合, 并结合预设 prompt 模板, 送入交互式语义匹配模型 T5 进行精排. 最终, T5 模型生成结果“是”和“否”, 作为精准排序模块的最终判断标准. 医学术语标准化总体流程图如图 1 所示.

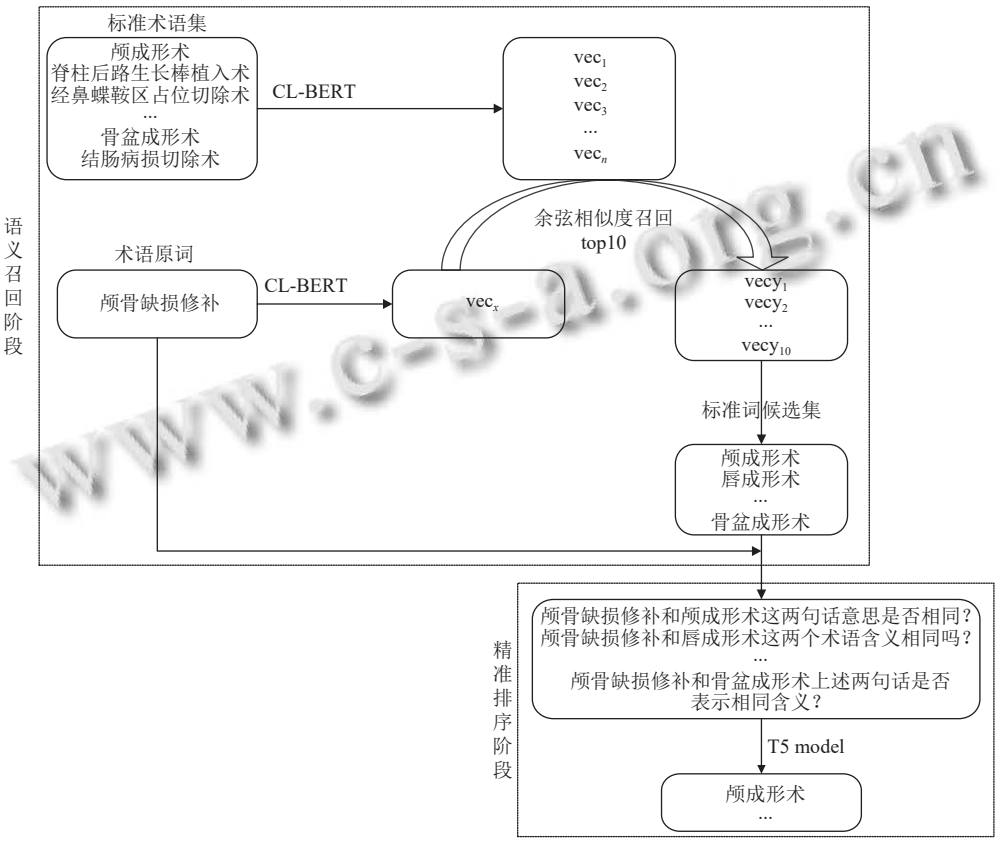

图 1 医学术语标准化总体流程图

2.1 语义召回阶段

语义召回模块是医学术语标准化任务的第一阶段, 在该阶段构建基于 RoBERTa-wwm 预训练模型[16]的双塔模型, 通过设计的有监督对比学习损失函数对双塔模型进行训练, 得到语义表征模型 CL-BERT. 通过 CL-BERT 模型分别生成原词语义向量和标准术语库中全部术语的语义向量, 根据向量之间的余弦相似度进行召回得到标准词候选集.

2.1.1 CL-BERT 语义表征模型

CL-BERT 使用基于 RoBERTa-wwm[16]的双塔模型结构, 左右两个塔共享权重参数. 双塔模型的整体结构如图 2 所示.

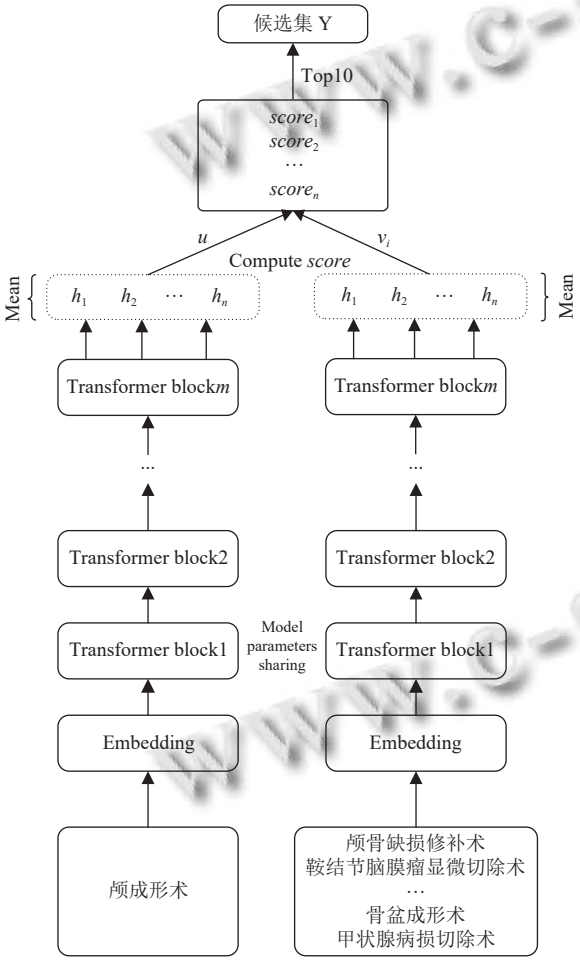

图 2 CL-BERT 整体结构图

首先将医学术语原词字符序列 $s=c_1c_2\cdots c_n$ 作为输入, 送入左塔的 Embedding 层中. 在 Embedding 层中, 将医学术语原词的字向量表示、句向量表示和位置向量表示进行叠加求和, 得到输出 $E=e_1e_2\cdots e_n$. 然后将 E 送入 Transformer block 层中, 每一层的 Transformer block 计算过程如式 (1), 式 (2):

$$H = LN(A^{i-1} + MHA(A^{i-1})) \tag{1}$$

$$A^i = LN(H + FFN(H)) \tag{2}$$

其中, $A^0=E$, A^i 表示第 i 层的 Transformer block 输出, MHA (multi-head attention) 为多头注意力机制, LN (layer normalization) 表示层归一化, FFN 为全连接神经网络.

通过公式计算得到最后一层 Transformer block 的输出为 $A^m=\{h_1, h_2, \cdots, h_n\}$, 对特征向量集合 A^m 在最后一个维度上纵向相加求和取平均得到 $u=\{x_1, x_2, \cdots, x_n\}$, 该向量 u 作为医学术语原词的语义向量表示.

接着, 将标准术语集 $G=\{g_1, g_2, \cdots, g_k\}$中的标准术语依次作为输入送入右塔, 将 g_i 送入右塔最终的输出结果 $v_i=\{y_1, y_2, \cdots, y_n\}$作为医学术语标准词 g_i 的语义向量表示. 获取到两个塔输出的医学术语语义向量 u, v_i 之后, 对语义向量 u, v_i 计算余弦相似度得到相似度得分, 计算公式为式 (3):

$$score(s, g_i) = \frac{u \times v_i}{\|u\| \times \|v_i\|} \tag{3}$$

通过式 (3) 计算原词和标准术语库中所有术语的相似度得分, 取得分最高的前 10 个医学术语标准词作为标准词候选集 Y, 计算公式为式 (4):

$$Y = Top_{10}(score(s, G)) \tag{4}$$

2.1.2 CL-BERT 模型训练

对于 CL-BERT 模型的训练, 首先按照相应的策略<原始词, 标准词, 1>和<原始词, 非标准词, 0>构造正负样本. 然后基于改进的有监督对比学习设计新型损失函数, 通过损失函数进行模型训练.

Khosla 等人将有监督的对比学习应用到图像领域. 对于有标签的样本, 标签相同的样本做正样本, 标签不同的做负样本, 基于对比学习构建损失函数, 缩小同类别样本的向量空间距离, 拉大不同类别样本的向量空间距离. 在 ImageNet 数据集上相比于传统有监督学习的交叉熵损失提升了 1% 达到了新的 SOTA[17].

本文借鉴对比学习的思想[18], 基于改进的有监督对比学习设计新型损失函数, 作为双塔模型训练过程中的损失函数. 该损失函数的总体思想为: 对于每一个 batch 的正负样本, 使正样本对的余弦相似度尽可能大于负样本对的余弦相似度, 使语义相近的文本在相似

度计算的过程中余弦值总大于语义不相近的文本, 而不将正负样本具体的余弦值当做模型训练过程中关注的重点. 损失函数的计算公式为式 (5):

$$f(loss)=\log\left(1+\sum_{(x,y)\in P,(m,n)\in N}\frac{e^{sim(um,vn)}/\tau}{e^{sim(ux,vy)}/\tau}\right) \quad (5)$$

其中, P 是训练集中标签为 1 的正样本, x 是原词, y 是标准词, ux、uy 是将 x、y 分别输入 CL-BERT 左塔和右塔得到语义向量表示. N 是训练集中的标签为 0 负样本, m 是原词, n 是非标准词, um、vn 是将 m、n 分别输入 CL-BERT 左塔和右塔得到语义向量表示. τ是自定义调节的温度系数, sim 表示余弦相似度.

由式 (5) 可见当分母 (正样本对的相似度) 越大于分子 (负样本对的相似度) 时, $\frac{e^{sim(um,vn)}/\tau}{e^{sim(ux,vy)}/\tau}$ 的值越趋于 0, 由于此处相似度计算使用的是余弦相似度, 该值趋于 0 而不等于 0. 由对数函数的变化规律可知 $\frac{e^{sim(um,vn)}/\tau}{e^{sim(ux,vy)}/\tau}$ 趋于 0 时损失值 $f(loss)$趋于 0. 上述整个过程的驱动因子是在一个 batch 中全部正样本对的余弦值之和更加大于全部负样本对的余弦值之和, 双塔模型在训练过程中也正是在不断调整权重参数以使其输出更加适应上述规律.

2.2 精准排序阶段

精准排序阶段主要包括 prompt 模板的构建和数据增强、T5 模型的构建和 FGM 对抗训练 3 个模块. 该阶段首先构建多样性的 prompt 模板, 把原词、标准词与多样性的 prompt 模板结合构成提问句, 将原始的 0、1 标签转变为“是”“否”回答句得到训练样本. 将提问句作为 T5 模型的输入, 回答句作为 T5 模型的输出训练 T5 模型, 此外, 加入对抗训练机制增强模型的鲁棒性.

2.2.1 prompt 模板的构建和数据增强

为了使预训练模型能够充分利用先验知识[19], 首先基于 prompt 思想构建多样化的模板, prompt 模板可以看作是一种检索方式, 用于从预训练语言模型中检索已经记忆的知识. prompt 模板如表 1 所示.

然后把原词、标准词与多样性的 prompt 模板, 通过“xxx 和 xxx+prompt 内容”的形式构成样本数据, 将原始的 0、1 标签转变为“是”“否”标签, 结合不同 prompt 模板数据增强后获得的样本数据如表 2 所示.

2.2.2 T5 模型

T5[20]作为典型的 Seq2Seq 模型, 结构上沿用了原始 Transformer[21]的 Encoder、Decoder 结构, T5 模型结构图如图 3 所示. T5 模型在 Transformer 的基础上做出如下一些改变: (1) 原始 Transformer 采用正弦余弦的计算方式获得位置编码[21]. T5 模型使用相对位置编码摆脱了传统绝对位置编码最大句子长度的限制[20,22]. (2) 去除了传统 Transformer 结构中 layer normalization 中的 bias, 将 layer normalization 放在残差连接的外面.

表 1 prompt 模板

序号	prompt内容
prompt 1	这两句话意思是否相同?
prompt 2	上述两句话是否表示相同含义?
prompt 3	这两个术语含义相同吗?
…	…
prompt *n*	上面这两句话意思一样吗?

表 2 结合 prompt 模板数据样本

标签	文本
是	“右中下肺叶切除术”和“肺叶切除术”两句话意思相同吗?
否	“右心导管术”和“经导管心脏微波消融术”这两句话是一个意思吗?
…	…
是	“颈管搔刮术”和“子宫颈管搔刮术”这两句话是在说同一件事吗?

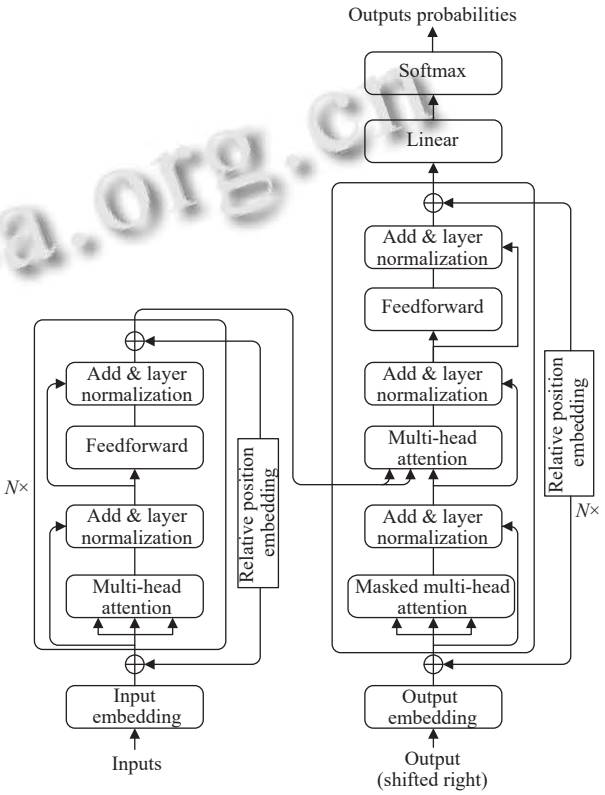

图 3 T5 模型结构图

使用上文通过不同 prompt 模板获得的提问句作为 T5 模型的输入, 经过 T5 模型生成答案“是”或“否”. 如图 4 所示左侧文本为 T5 模型 Encoder 端输入, 右侧文本为 T5 模型 Decoder 端的输出.

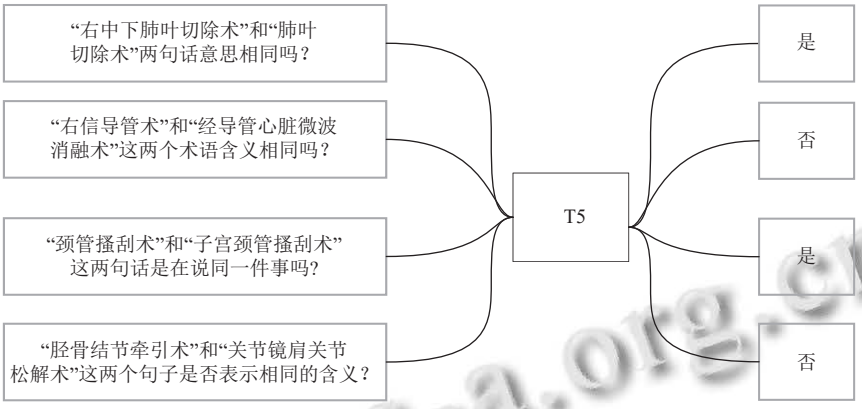

图 4 T5 模型的输入输出内容

2.2.3 FGM 对抗训练

为了提升模型的学习效果, 我们将对抗训练应用到对 T5 模型的训练过程中. NLP 中的对抗训练通过在模型的 embedding 层添加一些微小的扰动, 使其产生更具鲁棒性的向量, 同时对抗训练作为正则化的一种形式, 提升模型的泛化性[23]. 对抗训练的统一公式表示如式 (6):

$$\min \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in S} L(\theta, x+\delta, y)\right] \tag{6}$$

其中, $\mathcal{D}$代表输入样本分布, x 为输入, y 为标签, θ 为模型参数, S 代表扰动空间, L 代表损失函数.

在本模型中的对抗训练中主要步骤如下: Embedding 层的输出结果 x. 1) x 经过前向传播计算损失、反向传播计算得出梯度 r. 2) 将模型 Embedding 层梯度加到当前 Embedding 的输出 x 上, 得到新的输入 $x+r$. 3) 将 $x+r$ 前向传播和反向传播得到对抗梯度 t, 将对抗梯度 t 与步骤 1 中的梯度相加为 $t+r$. 4) 将 Embedding 恢复为 x 的值. 5) 根据步骤 3) 得到的梯度更新参数.

3 实验

3.1 数据集

在 ccks2019 医学术语标准化公开数据集上进行实验, ccks2019 数据集中所有的手术原词均来自于真实医疗数据, 并使用标准化术语库《ICD9-2017 协和临床版》进行标注, 训练集数据 6 000 条, 验证集数据 1 000 条. 原始数据集部分数据如表 3 所示.

为了提高模型对字面相似而语义不同的文本的区分能力, 使用最相似负样本召回的方式构建负样本, 字面相似度召回模型使用 BM25 模型. 在负样本构建过程中: 对于一对多的情况, 将标准词通过“##”拆分为多个标准术语后分别召回对应的相似文本, 去除正例之后召回标准术语总和为 15, 再从标准术语库中随机选择 10 个不包含正例的样本, 将上述方式获取的 25 个术语与原词分别配对构成 25 个负样本. 对于一对一或多对一的情况, 直接根据标准词召回前 15 个字面最相似的术语, 再从标准术语库中随机选择 10 个不包含正例的样本, 将上述方式获取的 25 个术语与原词分别配对构成 25 个负样本. 最后在训练过程中将正样本数量复制自增 25 倍, 以使正负样本保持均衡, 同时又保证负样本的多样性. 正负样本标签按照 0, 1 编码. 重构后的部分数据如表 4 所示.

表 3 原始数据集

原始词	标准词
右侧甲状腺叶切除术	单侧甲状腺叶切除术
双侧输尿管DJ管拔除术	单侧甲状腺叶切除术
支撑喉镜会厌囊肿切除术	内镜下会厌病损切除术
Mathieu尿道成形+阴茎下弯矫直术	尿道成形术##阴茎矫直术

表 4 处理后的数据集

原始词	标准词	标签
右侧甲状腺叶切除术	甲状旁腺切除术	0
右侧甲状腺叶切除术	单侧甲状腺叶切除术	1
右侧甲状腺叶切除术	甲状腺病损切除术	0
右侧甲状腺叶切除术	单侧甲状腺叶切除术	1

3.2 实验环境

本文实验环境配置如表 5 所示.

表 5 实验环境配置表

实验环境	参数
操作系统	Ubuntu 18.04.6
GPU	RTX3080Ti 12 GB×3
CPU	Intel E5-2680 v4 2.40 GHz
内存	256 GB
磁盘	8 TB
深度学习框架	Torch 1.7.0
cuda版本	11.4
Python版本	3.8.13

3.3 评价指标

斯皮尔曼相关系数是语义匹配模型的一个重要的衡量标准, 计算公式如式 (7) 所示, 其取值范围为 (−1, 1), ρ 越接近 1, 则模型的性能越好.

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{7}$$

其中, n 表示为测试集中样本对的个数, d_i 指的是 Xi 和 Yi 之间的等级差, Xi 和 Yi 的等级是指测试集中标签值和余弦相似度得分在其所在的列中从小到大排序后的位置.

3.4 实验过程及分析

在基于有监督对比学习的损失函数: $f(loss) = \log\left(1 + \sum_{(x,y)\in P,(m,n)\in N} \frac{e^{sim(um,vn)}/\tau}{e^{sim(ux,vy)}/\tau}\right)$ 中, 其中τ为可调节的温度系数, 表示对正负样本相似度差值的放大程度, 作为一个可调节的超参数, τ的取值对模型训练效果有直接的影响. 以 RoBERTa-wwm-ext-base 为基础预训练模型, 设置不同温度系数, 实验对比结果如表 6 所示.

表 6 不同温度系数取值实验对比

τ取值	斯皮尔曼相关系数
1/15	0.832 85
1/20	**0.843 32**
1/25	0.842 01
1/30	0.843 27

由表 6 可见τ取值 1/20 时, 本文设计的损失函数表现最好, 下文实验皆基于该系数取值.

实验中发现, 对于本文设计的损失函数, batch size 的大小对模型训练效果有影响, 不同 batch size 情况下实验效果对比如表 7 所示, 预训练模型 RoBERTa-wwm-ext-base, τ取值 1/20.

为了比较不同的语义表征模型效果的差别, 本文分别对比了 BERT、RoBERTa-wwm、MacBERT 这 3 种模型的 base 和 large 版本, 实验效果对比情况如表 8 所示, 参数固定为上文实验最佳参数. batch size=50, τ取值 1/20. 下文实验均采用该最佳参数.

表 7 不同 batch size 取值实验对比

batch size取值	斯皮尔曼相关系数
30	0.843 32
40	0.847 21
50	**0.851 13**
60	0.850 27

表 8 同类型模型召回排序效果对比

基础模型	斯皮尔曼相关系数
BERT-base-Chinese	0.832 85
RoBERTa-wwm-ext-base	**0.843 32**
MacBERT-base	0.842 01
BERT-large-Chinese	0.843 27
RoBERTa-wwm-ext-large	**0.858 73**
MacBERT-large	0.854 23

根据表 8 数据可见, RoBERTa-wwm-ext-large 在医学术语语义表征方面表现更加出色. 这主要是由于该模型采用了特殊的训练方式和大量的训练语料, 从而拥有更好的泛化能力和语义理解能力.

为了验证本文损失函数的有效性, 使用不同的基础预训练模型分别进行实验对比, 对比结果如表 9 所示, 评测指标均使用斯皮尔曼相关系数.

表 9 对比损失函数与交叉熵损失函数实验对比

基础模型	斯皮尔曼相关系数	
	交叉熵损失函数	对比损失函数
BERT-base-Chinese	0.843 28	**0.853 48**
RoBERTa-wwm-ext-base	0.843 11	**0.856 07**
MacBERT-base	0.844 77	**0.853 32**
ALBERT-base	0.836 33	**0.852 13**

由表 9 可见, 本文提出的基于有监督对比学习的损失函数, 在不同基础预训练模型上的表现相比于交叉熵损失函数均有所提升. 这是由于基于对比学习构建损失函数, 模型能更好地学习语义的相关性, 缩小了语义相近的术语向量空间距离, 拉大了语义不相近的术语向量空间距离, 提高了字向量对医学术语的表征能力.

在对 T5 模型进行 prompt tuning 阶段, 使用了对抗训练机制和多样化 prompt 模板进行数据增强. 为了验证对抗训练和数据增强的有效性进行了消融实验. 在实验中对比了不同的中文 T5 模型: T5-PEGASUS[24] 和 Mengzi-T5[25]. 实验结果如表 10 所示, 实验序号 1:

使用单一 prompt. 实验序号 2: 使用多种 prompt 数据增强. 实验序号 3: 加入 FGM 对抗训练. p: 精确率, r: 召回率, *F*1: 2(精确率×召回率)/(精确率+召回率).

表 10 消融实验

序号	T5-PEGASUS			Mengzi-T5		
	p	r	*F*1	p	r	*F*1
1	0.9801	0.9704	0.9752	0.9738	0.9725	0.9732
2	0.9834	0.9791	**0.9813**	0.9714	0.9823	0.9769
3	0.9781	0.9904	**0.9843**	0.9804	0.9797	**0.9801**

由表 10 序号 1 和序号 2、序号 3 分别进行对比可见, 使用多样化 prompt 数据增强和对抗训练对模型效果均有提升, 其中加入对抗训练机制提升效果较为明显. 对于中文 T5 预训练模型的选择, T5-PEGASUS 的效果比 Mengzi-T5 模型略有提升.

为了验证 T5+prompt tuning (T5 模型使用的 T5-PEGASUS) 方案的有效性, 将传统的交互式语义匹配模型 (分类模型) 与本文方案进行对比实验, 对比结果如表 11 所示.

表 11 T5+prompt tuning 对比实验

序号	基础模型	*F*1
1	CL-BERT+BERT-base-Chinese	0.8871
2	CL-BERT+RoBERTa-wwm-ext-base	0.9140
3	CL-BERT+MacBERT-base	0.8923
4	CL-BERT +ALBERT-base	0.9158
5	CL-BERT+T5+prompt tuning	**0.9206**

从实验数据可以看出通过 prompt 的引导, 对 T5 模型在预训练任务的基础上继续训练效果比传统的交互式语义匹配模型效果更好. 最终的实验结果也展示了本文提出的方法在医学术语标准化任务的可行性.

4 结论与展望

本文在解决医学术语标准化的问题上, 提出语义召回加精准排序的两阶段方案. 在语义召回阶段使用语义表征模型 CL-BERT, 分别得到原词和医学术语标准词的语义表征向量, 通过余弦相似度召回前 10 个最相似的标准术语得到标准词候选术语集. 在精准排序阶段利用 T5 结合 prompt tuning 构建语义精准匹配模型, 将原词和候选术语集分别进行精准排序, 得到最终的标准词. 在 1000 条验证集上两阶段串行结果 *F*1 值为 0.9206, 证明了本方法具有较高的精度, 为其他从事术语标准化研究者提供了参考. 下一阶段将尝试从医疗术语本身构成的特点入手, 例如加入“部位”“疾病性质”等特征, 进一步提升本文方法的预测准确率. 在预训练语言模型结构方面, 针对 BERT 模型词向量在高维空间呈现锥形分布的问题, 将结合语义匹配的任务特点对 BERT 模型结构进行改进, 并结合对比学习的思想重头预训练, 构建开箱即用的语义向量生成模型. 针对 prompt 模板不易选择, 寻找最佳 prompt 模板所需实验成本较高的问题, 将探索更加通用的 soft prompt 与 T5 模型结合, 进行交互式语义匹配.

参考文献

1 Ghiasvand O, Kate RJ. UWM: Disorder mention extraction from clinical text using CRFs and normalization using learned edit distance patterns. Proceedings of the 8th International Workshop on Semantic Evaluation. Dublin: Association for Computational Linguistics, 2014. 828–832. [doi: 10.3115/v1/S14-2147]

2 Kang N, Singh B, Afzal Z, *et al*. Using rule-based natural language processing to improve disease normalization in biomedical text. Journal of the American Medical Informatics Association, 2013, 20(5): 876–881. [doi: 10.1136/amiajnl-2012-001173]

3 Leaman R, Islamaj Doğan R, Lu ZY. DNorm: Disease name normalization with pairwise learning to rank. Bioinformatics, 2013, 29(22): 2909–2917. [doi: 10.1093/bioinformatics/btt474]

4 Li HD, Chen QC, Tang BZ, *et al*. CNN-based ranking for biomedical entity normalization. BMC Bioinformatics, 2017, 18(11): 385. [doi: 10.1186/s12859-017-1805-7]

5 Luo Y, Song GJ, Li PY, *et al*. Multi-task medical concept normalization using multi-view convolutional neural network. Proceedings of the 2018 AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 720.

6 Zhang YZ, Ma XJ, Song GJ. Chinese medical concept normalization by using text and comorbidity network embedding. Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM). Singapore: IEEE, 2018. 777–786. [doi: 10.1109/ICDM.2018.00093]

7 Tutubalina E, Miftahutdinov Z, Nikolenko S, *et al*. Medical concept normalization in social media posts with recurrent neural networks. Journal of Biomedical Informatics, 2018, 84: 93–102. [doi: 10.1016/j.jbi.2018.06.006]

8 赵兰枝, 史欣沅. 基于卷积神经网络的生物医学实体标准化研究. 科技创新与应用, 2022, 12(15): 30–35. [doi: 10.19981/j.CN23-1581/G3.2022.15.006]

9 Ji ZC, Wei Q, Xu H. BERT-based ranking for biomedical

entity normalization. AMIA Joint Summits on Translational Science Proceedings, 2020, 2020: 269–277.
10 Kalyan KS, Sangeetha S. Target concept guided medical concept normalization in noisy user-generated texts. Proceedings of the 2020 Deep Learning Inside Out (DeeLIO): The 1st Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. ACL, 2020. 64–73.
11 Chen LH, Varoquaux G, Suchanek FM. A lightweight neural model for biomedical entity linking. Proceedings of the 2021 AAAI Conference on Artificial Intelligence. AAAI, 2021. 12657–12665.
12 胡宇, 申德荣, 聂铁铮, 等. 面向生物医学实体链接的联合式学习方法. 计算机学报, 2022, 45(4): 748–765. [doi: 10.11897/SP.J.1016.2022.00748]
13 Li LQ, Zhai YK, Gao JH, *et al*. Stacking-BERT model for Chinese medical procedure entity normalization. Mathematical Biosciences and Engineering, 2023, 20(1): 1018–1036. [doi: 10.3934/MBE.2023047]
14 闫璟辉, 向露, 周玉, 等. 深度生成式模型在临床术语标准化中的应用. 中文信息学报, 2021, 35(5): 77–85. [doi: 10.3969/j.issn.1003-0077.2021.05.010]
15 韩振桥, 付立军, 刘俊明, 等. 结合 RoBERTa 与多策略召回的医学术语标准化. 计算机系统应用, 2022, 31(10): 245–253. [doi: 10.15888/j.cnki.csa.008757]
16 Cui YM, Che WX, Liu T, *et al*. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504–3514. [doi: 10.1109/TASLP.2021.3124365]
17 Khosla P, Teterwak P, Wang C, *et al*. Supervised contrastive learning. Proceedings of the 34th Advances in Neural Information Processing Systems. NeurIPS, 2020. 18661–18673.
18 张重生, 陈杰, 李岐龙, 等. 深度对比学习综述. 自动化学报, 2023, 49(1): 15–39. [doi: 10.16383/j.aas.c22042]
19 Gao TY, Fisch A, Chen DQ. Making pre-trained language models better few-shot learners. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL, 2021. 3816–3830.
20 Xue LT, Constant N, Roberts A, *et al*. mT5: A massively multilingual pre-trained text-to-text Transformer. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 483–498.
21 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.
22 Raffel C, Shazeer N, Roberts A, *et al*. Exploring the limits of transfer learning with a unified text-to-text Transformer. The Journal of Machine Learning Research, 2020, 21(1): 140.
23 Madry A, Makelov A, Schmidt L, *et al*. Towards deep learning models resistant to adversarial attacks. Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR, 2018.
24 Zhang JQ, Zhao Y, Saleh M, *et al*. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. Proceedings of the 37th International Conference on Machine Learning. ACM, 2020. 1051.
25 Zhang ZS, Zhang HQ, Chen KM, *et al*. Mengzi: Towards lightweight yet ingenious pre-trained models for Chinese. arXiv:2110.06696, 2021.

(校对责编: 孙君艳)