计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):245−253 [doi: 10.15888/j.cnki.csa.008757]

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

结合 RoBERTa 与多策略召回的医学术语标准化①

韩振桥[1,2], 付立军[1,3], 刘俊明[1,2], 郭宇捷[1,2], 唐珂轲[1,2,4], 梁　锐[4]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[2](中国科学院大学, 北京 100049)
[3](山东大学 大数据技术与认知智能实验室, 济南 250100)
[4](中康健康科技有限公司, 广州 510620)
通信作者: 付立军, E-mail: fu_lijun@ucas.ac.cn

摘　要: 针对传统的基于模板匹配、人工构建特征、语义匹配等解决术语标准化的方案, 往往会存在术语映射准确率不高, 难以对齐等问题. 本文结合医疗领域的文本中术语口语化、表达多样化的特点, 使用了多策略召回和蕴含语义评分排序模块来提升医学术语标准化效果. 在多策略召回模块中使用了基于 Jaccard 相关系数、TF-IDF、历史召回方法进行召回, 在蕴含语义评分模块使用了 RoBERTa-wwm-ext 作为判分语义模型. 首次在医学专业人员标注的基于 SNOMED CT 标准的中文数据集上验证了可用性. 实验证明, 在医疗知识特征的处理中, 本方法能够在医学术语标准化实际应用上达到不错的效果, 具有很好的泛化性及实用价值.

关键词: 术语标准化; 知识映射; 深度学习; RoBERTa-wwm-ext; SNOMED CT

引用格式: 韩振桥,付立军,刘俊明,郭宇捷,唐珂轲,梁锐.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. http://www.c-s-a.org.cn/1003-3254/8757.html

Combining RoBERTa with Multi-strategy Recall for Medical Terminology Normalization

HAN Zhen-Qiao[1,2], FU Li-Jun[1,3], LIU Jun-Ming[1,2], GUO Yu-Jie[1,2], TANG Ke-Ke[1,2,4], LIANG Rui[4]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Laboratory of Big Data and Artificial Intelligence Technolgy, Shandong University, Jinan 250100, China)
[4](Sinohealth Technology Limited, Guangzhou 510620, China)

Abstract: Traditional terminology standardization schemes based on template matching, artificially constructed features, semantic matching, etc., are often faced with problems such as low terminology mapping accuracy and difficult alignment. Given the colloquial and diverse expression of terminology in medical texts, modules of multi-strategy recall and implication semantic score ranking are used to improve the effect of medical terminology standardization. In the multi-strategy recall module, the recall method based on the Jaccard correlation coefficient, term frequency-inverse document frequency (TF-IDF), and historical recalls is employed. In the implication semantic scoring module, RoBERTa-wwm-ext is adopted as the scoring semantic model. The usability of the proposed method is validated for the first time on a Chinese dataset that is based on the systematized nomenclature of medicine-clinical terms (SNOMED CT) standard and annotated by medical professionals. Experiments show that in the processing of medical knowledge features, the proposed method can achieve favorable results in practical applications of medical terminology standardization and has high generalization and practical value.

Key words: term normalization; knowledge mapping; deep learning; RoBERTa-wwm-ext; systematized nomenclature of medicine-clinical terms (SNOMED CT)

① 基金项目: 国家社科基金 (21BTQ106)
收稿时间: 2022-01-14; 修改时间: 2022-02-15; 采用时间: 2022-03-11; csa 在线出版时间: 2022-07-07

1 引言

近年来, 随着国家对医疗健康的重视, 人们对自身健康的关注度也越来越高, 很多企业和研究单位都开始深入智能医疗与健康领域, 其中包括腾讯、阿里、京东、百度以及各 AI 医疗企业等, 共同推动了智能诊疗、医疗问答、临床辅助决策等技术的发展. 在医学场景中, 已经可以直观地感受到医学文本数量明显的增加, 其中医学文本包括医学文献、临床检测报告、电子病历记录、医疗保险记录等, 这些医学文本中包含了大量的可以挖掘利用的信息, 而这些数据中大多是非结构化或者半结构化, 如何更好地对这些数据进行有效的分析和利用, 是当前的研究热点和难点.

术语标准化能够帮助数据更合理的分析和利用并且提升下游任务的应用效果. 本文主要研究中文医疗文本的术语标准化. 医学术语标准化是将非正式的医学术语如“经皮髂骨成形术”, 映射到正式的医学概念, 如概念“骨盆成形术”, 然后再对应到相应的医学编码上. 这项任务在医学领域非常重要, 在临床上, 关于一种疾病、药品、症状等都有各种不同的写法 (包含非正式、非标准的形式还有误写等), 如果都能够归一到对应的术语上来, 它能够推动 AI 技术在医学应用系统上的落地, 如“CDSS (临床决策诊疗系统)”“DRGs (诊断相关分组管理系统)”等[1]. 并且这项技术在辅助诊疗、公共卫生检测、医疗检索等方面有巨大的作用.

在广大研究者的积极推动下, 关于术语标准化的研究经过了如下的几个阶段: 基于规则和字符词典匹配的方法[2,3]、基于机器学习的方法[4]、基于深度学习的方法[5,6]. 早期的基于规则的方法, 由于人工消耗较大且只能在特定的语料上达到满意的效果, 所以在处理比较复杂的数据时往往达不到预期. 后来随着机器学习、深度学习的发展, 人力构建规则的成本消耗得到很大的缓解, 相应术语标准化的准确率也获得了极大的提升. 由于深度学习方法的非线性建模能力更强、能够利用语义信息等优点, 所以在术语标准化的任务上的效果也能达到更好. 随着预训练模型 BERT[7] 的诞生, 因为它是通过未标注维基百科数据训练得到, 包含了丰富的先验知识和语义信息, 所以在术语标准化任务上利用预训练模型会比传统的深度神经网络如 LSTM[8] 有更优越的性能.

目前来说, 在使用 BERT 进行术语标准化任务时一般会采用的方法为直接排序和先召回再排序两种方式. 后一种方式能够相对减少排序时间的开销, 本次研究也是基于先召回再排序的思想.

本文在此基础上使用多策略召回排序的思路, 如图 1. 在第 1 阶段尽可能把正确的概念召回, 第 2 阶段使用蕴含语义评分模型将术语原词与候选概念进行语义相似度排序, 筛选出得分最高的概念. 同时, 之前的中文医疗领域术语标准化研究都是以 ICD9 或者 ICD10 (international classification of diseases, ICD) 为标准, 本文首次在 SNOMED CT (the systematized nomenclature of human and veterinary medicine clinical terms) 标准的数据上进行研究探索, 验证了本方法的有效性及使用 SNOMED CT 探索术语标准化的可能. 该实验结果证明, 本文提出的方法具有很强的实用性.

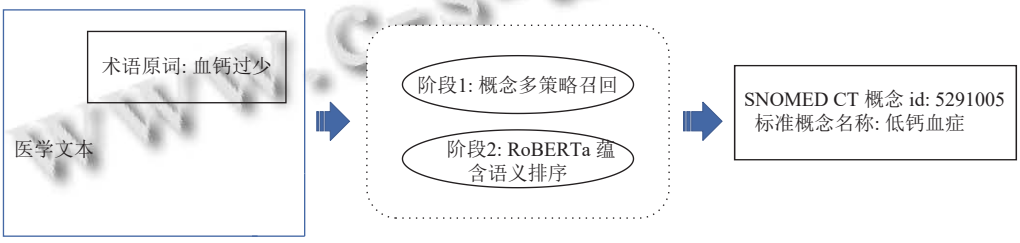

图 1 整体算法流程图

2 相关工作

本文主要的研究是在以中文为核心的医疗领域的术语标准化, 对于医学文本中的不规范的表达也就是术语原词, 经过标准化之后映射到正确的概念编码上来. 每一个标准的概念都有一个概念编码. 有表 1 不规范的表达, 最后经过术语标准化算法, 最终对应到正确的 SNOMED CT 编码.

医学术语标准化任务的目标是将医学文本中抽取的非标准的医学表达映射到正确的医学概念编码上, 以便于直接或给下游医学任务利用. 迄今为止, 已经积累了很多关于医疗领域的术语标准化研究工作.

早期的医学术语标准化工作主要是采用规则和机

器学习的方法. 文献 [9] 引入了一种新的基于编辑距离的方法来进行疾病名称标准化, 在 SemEval 2014 疾病名称标准化比赛[10] 中排名第二. 文献 [11] 使用了 5 种规则的 NLP 技术提升生物医学文本中的疾病名称标准化水平, 分别提升了 MetaMap[12] 和 Peregrine 系统[13] 的效果. 文献 [14] 提出一种多层筛选系统, 通过定义 10 种不同优先级的规则来度量术语原词和实体库中概念的相似性. 文献 [15] 利用线性模型对候选术语与概念名称之间的相似性进行评分, 并且运用了从训练集学习候选实体和概念名称的相似性的策略. 文献 [16] 在文献 [15] 的基础上采用了基于低秩矩阵近似的降维技术, 减少了参数量, 同时提升了疾病名称标准化在 NCBI 疾病语料上的效果.

表 1 术语原词-标准概念对应表

术语原词	标准概念	SNOMED CT ID
失血休克	失血性休克	355001
溃疡性口疮炎	溃疡性口炎	450005
巨唇(症)	巨唇症	643001
侵入肺曲霉病	侵袭型肺曲霉病	3214003
关节发炎	关节炎	3723001

近些年来, 随着深度学习的蓬勃发展, 来源于神经网络的深度学习技术在术语标准化任务上的表现不断突破, 这种不依赖于规则和人工特征的方案逐渐成为主流. 基于深度学习的术语标准化任务相关研究工作有: 文献 [17] 使用了不同语料训练的 Word2Vec[18] 语义向量表示, 并且利用卷积神经网络 (CNN) 和循环神经网络 (RNN) 提取特征, 大大超越了以 TF-IDF、BM25、向量相似度为基线的术语标准化水平. 文献 [19] 在文献 [18] 的基础上增加了医疗健康相关的文本训练获得医疗专业的词嵌入 (word embedding), 从而更好地表示医学概念的语义特征, 在数据集上获得新的 SOAT (state of the art). 文献 [20] 提出了一种疾病名称和术式名称结合的多任务医学术语标准化框架, 利用多视角 CNN 提取特征, 并且对两个任务引入权重共享层, 利用疾病名称和术式名称之间的相关性更好地进行术语标准化. 文献 [21] 针对术语标准化任务构建了端到端的模型结构, 运用结合 attention[22] 的双向 LSTM 和 GRU 结构[23] 提取候选实体特征, 与 UMLS 系统中的标准词概念特征拼接, 运用 Softmax 函数进行评分, 证明了比单纯使用 CNN 结构进行术语标准化有更好的效果. 文献 [24] 考虑到标注医疗数据需要丰富的专业知识和时间开销, 提出了一种利用共病网络 embedding 的疾病名称标准化的无监督方法, 接近了经典有监督学习的准确性.

但是以上方法都存在一个问题: 初始的词嵌入并不能表示一词多义, 词向量的特征包含不够丰富, 随着预训练模型的提出, 在术语标准化任务上有了如下的研究.

文献 [25] 采用基于字符级 ELMo 向量[26] 与传统的 Word2Vec 词向量拼接共同表示最终的词向量的, 以获得包含更丰富信息的词向量. 并利用 BiLSTM 提取候选实体的特征, 最终结果超越了以 BIGRU-attention 进行术语标准化的 SOAT. 文献 [27] 将标准化任务视为一个分类问题, 在 3 个不同的数据集上进行实验, 对医学概念标准化任务的模型进行细粒度的评估. 通过 BERT、ELMO、RNNs 模型进行语义表示, 分别对比在术语标准化上的效果, 得出 BERT 在医学概念标准化上有更好的效果、神经网络的结构会影响术医学概念标准化的准确性等结论. 文献 [28] 比较 BERT/BioBERT/ClinicalBERT 在生物医学实体标准化任务上的准确性, 结论得出对预训练模型进行微调可以显著提升生物医学实体标准化水平. 文献 [29] 提出了一种生成和排序的框架解决医学术语标准化问题. 第一阶段使用 Lucene 工具生成候选对象, 之后使用 BERT 进行候选实体打分.

基于规则的方法需要根据不同的场景设定不同的规则, 费时费力同时可移植性不强. 基于机器学习的方法虽然在一定程度上缓解了人工消耗, 但由于缺乏语义信息的局限性且不能考虑上下文信息, 它不能在更为复杂的医学术语标准化任务上表现得很好. 深度学习在文本建模上具有强大的表征能力, 不仅可以更好地表示词语和文本, 还可以学习到词语的上下文关系和重要词语的信息[30]. 随着预训练语言模型的诞生, 因为其在上下文中可以获得更为丰富的语义特征, 且使用基于预训练语言模型的方法在很多自然语言处理任务上都达到了最好的水平, 所以现在利用预训练模型模块实现医学术语标准化任务也成为了主流.

本文的研究也是基于预训练模型提高医学术语标准化任务准确率. 由于在医学术语标准化第一阶段的召回过程中单一的方法往往不能够覆盖大部分正确概念, 为此本文提出了多策略召回的方案, 极大提升了第 1 阶段正确概念的召回率. 结合第 2 阶段使用 RoBERTa-WWM-ext[31] 进行蕴含语义排序, 术语标准化最终的准

确性得到有效提高.

3 模型介绍

3.1 问题定义

在基于 SNOMED CT 标注的术语标准化数据集中, 设标准概念数量为m, 其中概念集为$C=\{c_1, c_2, \cdots, c_m\}$.

术语原词为t, 经过第 1 阶段混合召回, 将概念集缩小到$G=\{g_1, g_2, \cdots, g_k\}$, 其中$k<m$, 在第 2 阶段经过精细化排序, 在候选概念$G$中选择一个得分最佳的概念, 作为最终术语标准化的结果. 两阶段实现术语标准化, 其核心在于要在召回阶段能够尽量地把正确概念召回, 召回的概念作为候选实体, 这决定了后续排序阶段效果的上限. 在排序阶段, 要能够精细化排序得出最佳的概念.

3.2 构建两阶段术语标准化模型

本文提出的模型, 总体分为两部分, 下面会对这两部分分别介绍.

第 1 部分是多策略召回阶段, 多策略召回分为 3 个小模块. 通过计算术语原词与术语库中所有概念的 Jaccard 相关系数, 取 Jaccard 相关系数最高的作为候选实体的一部分. 同时也在所有的概念经过分词之后训练一个 TF-IDF 模型, 这样就能获得所有的分词权重, 之后把术语原词作为一个 query, 计算 query 与所有概念的相关性, 取相关性最高的概念作为候选实体. 同时结合历史召回方法对候选实体进行召回.

第 2 部分是蕴含语义排序模块, 使用了 RoBERTa-wwm-ext 模型, 计算术语原词与候选实体的语义相似性蕴含分数, 再进行蕴含分数排序, 选取得分最高概念, 图 2 展示了整体的模型架构图.

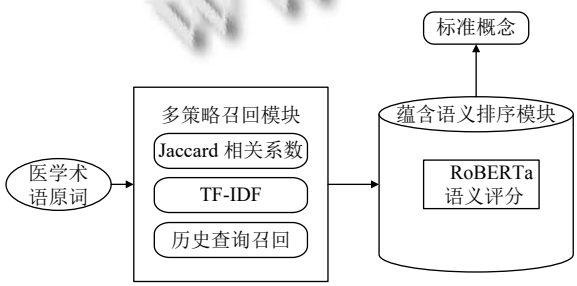

图 2 整体模型框架图

3.2.1 候选概念多策略召回模块

这个召回模块主要由 3 个小部分组成: 历史召回、TF-IDF 相关性召回、Jaccard 相关系数召回模块, 下面会介绍各小部分的召回原理, 图 3 展示了多策略召回的具体流程.

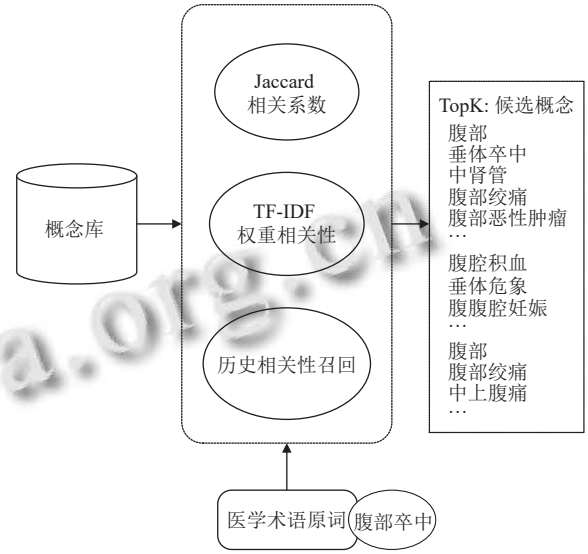

图 3 候选概念多策略召回

TF-IDF 召回原理: TF-IDF 是一种高效的计算特征权重的算法, 其可以用来解决短文本的相似度的问题. 本文将它用来作为术语原词和概念库中的概念相关性比较的算法, 通过此算法将术语原词匹配到部分最相关的概念.

首先要计算概念库中特征词的权重, 将概念库中的每个概念进行分词, 对于概念库中概念中的特征词其特征权重的计算公式如式 (1):

$$\begin{aligned} TF\text{-}IDF(C_i, w_j) &= TF(C_i, w_j) \times IDF(C_i, w_j) \\ &= \frac{N(w_j)}{n} \times \log \frac{m}{M(w_j)+1} \end{aligned} \tag{1}$$

其中, $N(w_j)$是w_j在C_i中出现的次数; m是概念库中的概念总数; $M(w_j)$是概念库集和中含有w_i的概念数.

接下来计算术语原词与概念库中每个概念的相关性得分, 对术语原词s进行分词产生词语列表$[v]$, C_i产生的分词列表$[w]$, 计算s和C_i的相关性得分如式 (2):

$$\begin{aligned} score(s, C_i) &= score([v], [w]) \\ &= \sum TF\text{-}IDF(C_i, w_j),\ w_j \in [v], w_j \in [w] \end{aligned} \tag{2}$$

Jaccard 相关系数召回原理: Jaccard 相关系数主要计算符号度量的个体之间的相似程度. 对于术语原词s_1和概念s_2, 要计算他们的 Jaccard 相关系数, 可以先将s_1、s_2分别分为字符集合 A 和字符集合 B. 他们的 Jaccard

相关系数计算如式 (3):

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \tag{3}$$

历史召回原理: 历史召回的算法也是采用的 TF-IDF, 它是对训练数据中的术语原词进行召回而不是直接对概念库中的标准词进行召回, 其优势在于能更加充分的利用训练集中的信息. 它直接对训练集中所有的术语原词进行计算获得特征分词的权重, 之后计算待标准化术语与训练集中术语原词的相关性, 最后取相关性最高的 top-k 个术语原词对应的标准概念作为候选概念集. 图 4 展示了历史召回的算法原理.

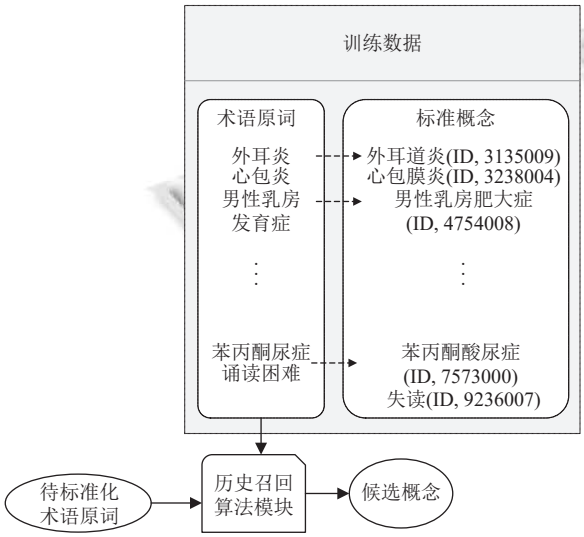

图 4 历史召回算法流程图

训练集中数据是以“术语原词-概念”的形式出现, 所以找到最匹配的术语原词, 也就间接地找到了需要的概念, 这种方法能够在召回阶段更充分利用数据.

3.2.2 蕴含语义排序模块

候选实体排序阶段, 要对第 1 阶段召回的所有概念进行打分排序, 这样才能选择与术语原词最对应的概念作为答案. 候选实体排序模块由 RoBERTa-wwm-ext 蕴含语义相似性评分模块构成.

RoBERTa-wwm-ext 作为蕴含分数计算模型, 主要是通过训练一个二分类模型, 之后预测原词和术语库中的概念之间的蕴含分数 (其中把预测为 1 的概率作为蕴含分数), 对术语库中的每个概念进行评分. RoBERTa-wwm-ext 是在 BERT 的基础上进行的改进, 以下主要介绍二者的区别.

BERT 采用预训练-微调的模式, 问世以来在解决实体识别、文本分类、自然语言推断等多个自然语言处理获得了 SOTA, 给学术界提供了很多的参考.

BERT 的全称是 (bidirectional encoder representation from Transformers), 本文中作为语义排序的基础模型. 在 BERT 之前预训练语言模型有 ELMo (embedding from language model) 和 GPT (generative pre-training), 但是这两种方法都只采用了一个预训练目标, 而且没有充分的利用上下文信息. BERT 采用 Transformer 的 encoder 作为基本组成, 能够充分结合上下文本信息进行有效的训练. 与早期提出的训练语言模型的目标“预测下一个词”不同的地方在于 BERT在单词级别和句子级别设置了两个目标: 掩码语言模型 (masked language model, MLM) 与预测下一句 (next sentence predict, NSP) 模型. 其中 MLM 可以理解为完型填空做法的思路, 模型随机会 mask 每个句子中 15% 的词, 利用上下文信息来预测这些词. MLM 具体做法是 80% 的词用 [mask] 替换原来的词, 10% 的词随机取一个词替代 mask 的词, 10% 词保持不变. 预测下一句训练过程的具体做法是选取一些句对与, 其中 50% 的数据是的下一句, 剩余 50% 是从语料库中随机选择, 通过对句对进行二分类训练来学习句子间的关系. 通过这两个目标训练出的 BERT 模型, 具有很强的字词级别的表征能力.

RoBERTa-wwm-ext 与 BERT 模型的基本结构基本相同, 改进更多的是从训练集和训练策略角度来提升, 主要有以下几点: 首先相对于 BERT 的静态掩码机制采取了动态掩码机制, 在 BERT 中训练数据时, 一条样本只进行一次随机 mask, 在训练时 mask 的位置都保持不变, 动态 mask 在每次训练前会动态生成一次 mask, 这种方法提高了模型输入的随机性, 使模型可以学习更多的句式. 另外它使用了更大的 batch size 进行训练, 被实验验证有更好的效果. 同时采用字节对编码 (BPE) 进行文本数据处理, 使用了更多的数据同时进行训练, 且取消了 NSP 任务, 提升了效率. 且采用了 WWM (全词掩码) 策略, 相较于 BERT 的单字掩码, 先进行分词, 如果有词中的部分字符被 mask, 那么整个词都将会被 mask, 这样做 RoBERTa-wwm-ext 能够更好地学习词级别的信息[31–33].

本文使用 RoBERTa-wwm-ext 模型将语义间相关性判别转化为一个二分类模型, 图 5 展示了 RoBERTa-wwm-ext 模型结构作为语义蕴含评分模块.

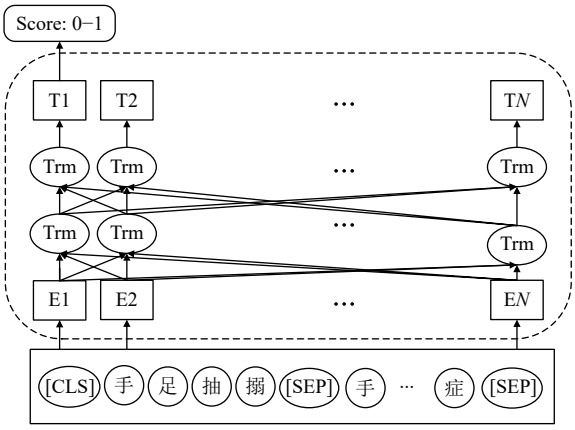

图 5 RoBERTa-wwm-ext 语义蕴含评分模型结构

分别按照相应的策略构造<术语原词, 概念, 1>和<术语原词, 非正确概念, 0>的正负样本, 用此数据来训练蕴含语义评分模型. 将样本输入到模型中具体格式为“[CLS] 手足抽搐 [SEP] 手足抽搐症 [SEP]”, 经过 RoBERTa-wwm-ext 编码获得 [CLS] 的隐藏层状态表示, 接着输入到全连接层, 经过 Softmax 函数打分, 其中 [CLS] 和 [SEP] 分别表示用于分类的令牌、分隔术语原词和候选概念的令牌.

蕴含语义评分模型使用 Softmax 作为分类回归函数, 模型采用交叉熵损失函数进行优化, 使用类别为 1 的概率作为语义蕴含分数. 其中蕴含分数及最后预测对应概念的计算过程如式 (4)–式 (5):

$$score(s,g_i)=\underset{l=1}{Softmax}(FFNN(RoBERTa\text{-}wwm\text{-}ext(E1,\cdots,EN))) \quad (4)$$

$$y=\arg\max(score(s,G)) \quad (5)$$

其中, $FFNN$ 表示前馈神经网络层, l为类别标签, y表示最终对应的概念.

本文需要比较的是术语原词和概念库当中概念的语义相似性, 虽然使用的是二分类的方法, 但是需要二分类模型评出概念相似程度的高低, 涉及到排序, 所以其精度要求更高, 难度上比传统二分类任务的 0.5 作为阈值更难. 所以对构造二分类模型的训练样本一定要经过特殊处理, 这样模型才能够更好地学习到语义的相关性.

方法 1. 构造语义模型数据集方法

1) 构建空的数据列表 datas=[], 将其作为语义模型训练需要的数据;
2) 设置困难负样本数量 k, 随机负样本数量 m, 正样本数量 n;
3) 训练集术语原词使用 Jaccard 相关系数召回概念库中得分最高的前 k 个负样本, 将 k 个样本处理为<org, neg, 0>并入 datas;
4) 训练集术语原词从概念库中随机抽取 m 个负样本, 将 m 个样本处理为<org, neg, 0>并入 datas;
5) 训练集术语原词重复 n 条作为正样本, 将 n 个样本处理为<org, neg, 1>并入 datas;
6) 对 datas 进行随机打乱;

4 实验过程与结果评估

4.1 实验数据

SNOMED CT 是目前国际上认可的且比较全面的医学术语集, 其内容包括了临床所需的基本信息. SNOMED CT 的概念表收录了大量具有唯一含义并经过逻辑定义的概念, 分类编入 18 个顶级概念轴 (hierarchy) 中, 分别包括临床发现、操作/介入、身体结构等[34].

本次实验的数据是以 SNOMED CT 为标准, 医疗相关人员进行标注, 获得“术语原词-概念”标注数据 9 000 余条, 此次标注中选取 SNOMED 术语库中在医学文本中常见的概念 15 001 个 SNOMED CT 概念作为概念标准, 这些数据均来自现实的医学场景. 按照比例 6:2:2 划分为训练集、开发集、测试集, 数据集中的真实数据如表 2.

表 2 数据集中的数据形式

原始词	标准词	SNOMED CT ID	SNOMED CT FSN
毛细支气管炎	小支气管炎	4120002	Bronchiolitis
溃疡性口疮炎	溃疡性口炎	450005	Ulcerative stomatitis
骨软化病	骨质软化症	4598005	Osteomalacia
血磷酸盐过少	低磷血症	4996001	Hypophosphatemia

数据集中最长的术语原词为“库兴氏综合征 (由于各种原因引起的肾上腺糖皮质激素慢性分泌过多, 表现为肥胖伴有高血压等一系列症状)”, 最短的术语原词为“腿”.

4.2 评价指标

4.2.1 总项评价指标

对于分类任务, 总体的评价指标有如下几个标准: 准确率、精确率、召回率、$F1$ 值, 其计算方式如式 (6)–式 (9):

$$accuracy=\frac{TS}{N} \quad (6)$$

$$precision=\frac{TP}{TP+FP} \quad (7)$$

$$recall=\frac{TP}{TP+FN} \quad (8)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{9}$$

式 (6) 中, TS代表被预测正确概念的样本数, N代表样本总量, 本文最终目标使用$accuracy$作为评价指标.

4.2.2 分项评价指标

本文是术语标准化任务作为候选概念召回和候选概念排序两阶段任务来实现的, 所以在这两个分项中也有不同的指标来衡量他们的效果. 候选概念召回阶段使用召回率 ($recall$) 作为评价指标. 其中$recall$的计算公式如式 (10):

$$recall = \frac{\sum_{i=1}^{n}(G_i \cap T_i)}{N} \tag{10}$$

其中, G_i代表第数据集中第i个术语原词召回来的候选概念集, T_i表示第i个术语原词的正确答案, N为术语原词总数.

4.3 实验过程及分析

4.3.1 参数设置

本文中使用了 RoBERTa-wwm-ext 作为蕴含语义评分模型, 使用了 Adam 作为优化器, 实验采用内存大小为 11 GB, 一张 2080ti GPU 显卡. Dropout rate 设置为 0.1, 学习率为 2E–5, batch_size 为 64, 隐藏层维度为 768, 最大的句子长度为 64.

4.3.2 结果分析

在第一阶段召回主要做实验对比单召回方式, 以及本文中多策略召回的效果, 表 3 展示了不同方法的候选概念召回率.

表 3 候选概念召回率

召回方式	Top-1	Top-3	Top-5	Top-10	Top-20
Jaccard系数召回	0.525	0.707	0.762	0.820	0.866
历史召回	0.284	0.426	0.470	0.515	0.548
TF-IDF召回	0.285	0.438	0.487	0.549	0.605
TF-IDF + 历史召回	0.457	0.621	0.670	0.718	0.756
TF-IDF + 历史 + Jaccard系数召回	0.843	0.872	0.919	0.983	0.985

表 3 的结果来看, 本文提出的召回策略具有明显优势. 单一方法的召回都不能够完全覆盖正确的概念. 结合多策略召回, 其中每个召回方式都取前 10 最高得分能够达到 0.983 的召回率, 基本能够覆盖正确答案, 所以将它作为蕴含语义排序模型的输入部分. 在候选概念蕴含语义排序阶段, 为了获得更强的语义排序模型, 本文采用了不同的策略构建负样本. 在训练语料中引入了困难负样本, 再结合不同的训练样本比例, 得到当前效果最佳的蕴含语义模型训练方案. 不同构造样本的策略效果如表 4.

表 4 样本构建对结果影响

序号	(P, N_{random}, N_{hard})	准确率
1	(1, 1, 0)	0.409
2	(1, 1, 1)	0.332
3	(5, 5, 0)	0.610
4	(5, 5, 5)	0.653
5	(10, 10, 0)	0.621
6	(10, 10, 10)	0.796
7	(20, 20, 0)	0.638
8	(20, 20, 20)	**0.878**
9	(25, 25, 0)	0.604
10	(25, 25, 25)	0.843

表 4 中, P 代表训练集中的正样本重复次数, N_{random} 数代表随机负样本数量, N_{hard} 表代表困难负样本数量. 为了维持训练集正负样本的比例, 所以始终将训练正负样本维持在 1:1 与 1:2. 从蕴含语义排序的结果来看, 在一定范围内引入困难负样本会显著的提升蕴含语义排序模型的效果并且增加正、负样本的数量也可以提升模型的效果. 为了比较不同的语义模型蕴含排序效果的差别, 本文分别在同结构的网络模型进行了对比实验. 在召回策略相同、正负样本构造分别是 (P, N_{random}, N_{hard})=(20, 20, 20) 的条件下, 各语义蕴含模型的效果如表 5.

表 5 部分同类型模型蕴含语义排序效果对比

序号	方法	准确率
1	多策略+BERT-base	0.869
2	多策略+BERT-wwm-ext	0.872
3	多策略+RoBERTa-base	0.864
4	本文方法	**0.878**

从结果来看, 在使用相同的召回策略且蕴含语义模型的正负样本构建策略均相同的情况下, 使用 RoBERTa-wwm-ext 作为蕴含语义排序模型展现了它有更强的蕴含语义表征能力, 并且效果比同类其他模型的效果更好. 这是由于其模型的训练方式和丰富的训练语料带来的优势, 实验结果也展现了本方法在以 SNOMED CT 为标准的医学术语标准化上的可行性及优越性.

5 结论与展望

本文在解决医学术语标准化的问题上, 提出了一

种结合 RoBERTa 与多策略召回的方法, 该模型使用 RoBERTa-wwm-ext 作为蕴含语义排序模型. 首次在医疗标准 SNOMED CT 标注的数据上进行实验验证, 证明了本方法的有效性, 为其他从事 SNOMED CT 标准进行的术语标准化工作者提供了参考. 本文将医学术语标准化分为两个阶段来执行, 第 1 阶段是多策略召回, 第 2 阶段是蕴含语义排序. 在多策略召回阶段, 由于医学术语的表达多样化与口语化的特点, 往往通过一种召回方法召回候选概念效果欠佳, 而本文提出的多策略的召回方法可以召回 98.3% 的正确候选概念. 在蕴含语义排序阶段, 为了构建强大语义模型, 本文引入了困难负样本进行训练, 并且构造不同数量的正负样本比例确定蕴含语义模型的训练方式, 最终蕴含语义排序模型的效果得到极大提升. 通过对比本文模型和其他同类型模型的基于 SNOMED CT 医学术语标准化效果, 本文提出的模型有更高的准确性, 准确率达 87.8%.

由于医疗领域的术语一字之差可能完全表达的是两个不同的概念、术语原词与概念之间没有交集等情况. 在以后的工作中希望能够引入外部信息, 或者根据医疗数据特点引入特征词典来提高医学术语标准化的水平, 同时也希望对一个术语对应多个概念或者多个术语对应多个概念的方向去展开研究.

参考文献

1 陈漠沙, 仇伟, 谭传奇. 基于 BERT 的手术名称标准化重排序算法. 中文信息学报, 2021, 35(3): 88–93. [doi: 10.3969/j.issn.1003-0077.2021.03.008]

2 Tsuruoka Y, McNaught J, Tsujii J, *et al*. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics, 2007, 23(20): 2768–2774. [doi: 10.1093/bioinformatics/btm393]

3 Lee HC, Hsu YY, Kao HY. AuDis: An automatic CRF-enhanced disease normalization in biomedical text. Database, 2016, 2016: baw091. [doi: 10.1093/database/baw091]

4 Xu J, Lee HJ, Ji ZC, *et al*. UTH_CCB system for adverse drug reaction extraction from drug labels at TAC-ADR 2017. Proceedings of TAC 2017. Gaithersburg: TAC, 2017. 1–6.

5 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]

6 Luo YF, Sun WY, Rumshisky A. A hybrid normalization method for medical concepts in clinical narrative using semantic matching. AMIA Joint Summits on Translational Science Proceedings, 2019, 2019: 732–740.

7 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2018. 4171–4186.

8 Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. Proceedings of the 13th Annual Conference of the International Speech Communication Association. Portland: ISCA, 2012. 194–197.

9 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.

10 Pradhan S, Elhadad N, Chapman WW, *et al*. SemEval-2014 task 7: Analysis of clinical text. Proceedings of the 8th International Workshop on Semantic Evaluation. Dublin: Association for Computational Linguistics, 2014. 54–62.

11 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]

12 Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. Proceedings of AMIA 2001, American Medical Informatics Association Annual Symposium. Washington: AMIA, 2001. 17–21.

13 Schuemie MJ, Jelier R, Kors JA. Peregrine: Lightweight gene name normalization by dictionary lookup. Proceedings of the BioCreative 2 Workshop. Madrid, 2007: 131–140.

14 D'Souza J, Ng V. Sieve-based entity linking for the biomedical domain. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015. 297–302.

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

16 Leaman R, Lu ZY. Automated disease normalization with low rank approximations. Proceedings of BioNLP 2014. Baltimore: Association for Computational Linguistics, 2014.

24–28.
17 Limsopatham N, Collier N. Normalising medical concepts in social media texts by learning semantic representation. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 1014–1023.
18 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv: 1301.3781, 2013.
19 Lee K, Hasan SA, Farri O, *et al*. Medical concept normalization for online user-generated texts. Proceedings of 2017 IEEE International Conference on Healthcare Informatics (ICHI). Park City: IEEE, 2017. 462–469. [doi: 10.1109/ICHI.2017.59]
20 Luo Y, Song GJ, Li PY, *et al*. Multi-task medical concept normalization using multi-view convolutional neural network. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 5868–5875.
21 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]
22 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473, 2014.
23 Chung J, Gulcehre C, Cho KH, *et al*. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv: 1412.3555, 2014.
24 Zhang YZ, Ma XJ, Song GJ. Chinese medical concept normalization by using text and comorbidity network embedding. Proceedings of 2018 IEEE International Conference on Data Mining (ICDM). Singapore: IEEE, 2018. 777–786. [doi: 10.1109/ICDM.2018.00093]
25 Subramanyam KK, Sangeetha S. Deep contextualized medical concept normalization in social media text. Procedia Computer Science, 2020, 171: 1353–1362. [doi: 10.1016/j.procs.2020.04.145]
26 Peters ME, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018. 2227–2237.
27 Miftahutdinov Z, Tutubalina E. Deep neural models for medical concept normalization in user-generated texts. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence: Association for Computational Linguistics, 2019. 393–399.
28 Ji ZC, Wei Q, Xu H. BERT-based ranking for biomedical entity normalization. AMIA Joint Summits on Translational Science Proceedings, 2020, 2020. 269–277.
29 Xu DF, Zhang ZY, Bethard S. A generate-and-rank framework with semantic type regularization for biomedical concept normalization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. Online: ACL. 2020. 8452–8464.
30 孙曰君, 刘智强, 杨志豪, 等. 基于 BERT 的临床术语标准化. 中文信息学报, 2021, 35(4): 75–82. [doi: 10.3969/j.issn.1003-0077.2021.04.011]
31 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]
32 Liu Y, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv: 1907.11692, 2019.
33 王曙燕, 原柯. 基于 RoBERTa-WWM 的大学生论坛情感分析方法. 计算机工程, 2021: 1–9. (2021-10-18)[2022-01-13]. [doi: 10.19678/j.issn.1000-3428.0062008]
34 郭玉峰, 刘保延, 崔蒙, 等. SNOMED CT 内容简介. 中国中医药信息杂志, 2006, 13(7): 100–102. [doi: 10.3969/j.issn.1005-5304.2006.07.058]

(校对责编: 孙君艳)