文章编号:2096-1472(2023)11-0039-04
DOI:10.19644/j.cnki.issn2096-1472.2023.011.009

基于 PubMedBERT 预训练模型的医学术语对齐方法研究

王怡茹，郑建立，周浩然

(上海理工大学健康科学与工程学院，上海 200093)

✉ e_wangyiru@163.com; zhengjianli163@163.com; zhouhaoran1908@163.com

摘 要:随着互联网大健康数字化时代的到来,健康数据海量增长,为解决医疗数据集成应用中异构数据的术语标准化问题,提出一种利用 PubMedBERT 计算语义相似度实现医学术语对齐的技术。使用特定医学领域预训练模型,结合缩略词扩展方法增强语义信息,并与传统相似度计算模型、BERT(Bidirectional Encoder Representations from Transformers)及其变体相比较。在测试语料上的实验表明,缩略词扩展后 PubMedBERT 预训练模型 TOP1 的准确率提高了 18.79%,PubMedBERT 模型 TOP1、TOP3、TOP5、TOP10 的准确率分别达到 78.49%、85.69%、87.44%、89.54%,优于其他对比模型。该方法可以为医学术语对齐工作提供一种智能化的解决方案。

关键词:语义相似度;术语对齐;缩略词扩展;PubMedBERT

中图分类号:TP391.1 **文献标志码**:A

Research on Medical Term Alignment Method Based on PubMedBERT Pre-training Model

WANG Yiru, ZHENG Jianli, ZHOU Haoran

(*School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai* 200093, *China*)

✉ e_wangyiru@163.com; zhengjianli163@163.com; zhouhaoran1908@163.com

Abstract: In the context of digital era of Internet health, there is a massive growth of health data. In order to solve the problem of terminology standardization for heterogeneous data in medical data integration applications, a technology using PubMedBERT to calculate semantic similarity to achieve medical terminology alignment is proposed. This technology uses pre-trained models in specific medical fields, and enhances semantic information with abbreviation expansion methods. Then it is compared with traditional similarity calculation models, BERT (Bidirectional Encoder Representations from Transformers), and their variants. The experiment on the test corpus shows that the accuracy of PubMedBERT pre-trained model TOP1 has improved by 18.79% after abbreviation expansion, and the accuracy of PubMedBERT models TOP1, TOP3, TOP5, and TOP10 reaches 78.49%, 85.69%, 87.44%, and 89.54%, respectively, which is superior to other comparative models. This method can provide an intelligent solution for medical terminology alignment work.

Key words: semantic similarity; term alignment; expansion of abbreviations; PubMedBERT

0 引言(Introduction)

随着健康医疗数据资源的快速增长和网络信息技术的蓬勃发展,各种电子化健康信息数据分散在不同的系统和应用中。对于同一医学术语,不同的医疗卫生机构有不同的表达方式,医疗领域的数据共享性低、缺乏语义互操作性[1]。医学数据的处理普遍依赖于专业人员手动查找和对齐,映射周期长、规范程度低,已成为医学数据集成和再利用的主要瓶颈。标准化医学术语作为医学概念的形式化表示,是解决语义互操作性问题的基础,术语对齐是提高异构数据语义互操作性的重要途径[2]。因此,探索一种有效的医学术语对齐方法有利于促进卫生保健领域的数据共享和临床信息的电子交换。

收稿日期:2023-03-17

1 相关研究(Related research)

20世纪50年代,国际医学术语标准及体系初步建立。世界卫生组织、医疗信息标准委员会、美国病理学会、国际卫生术语标准开发组织等长期致力于医学术语相关标准的制定、发布和更新工作。如今,医学术语标准化工具的发展逐渐趋于稳定,成果颇丰。

针对国际标准医学术语集同本地医学术语的对齐工作已有许多的研究。例如,赵云松等[3]尝试利用R语言相关技术的扩展包,建立体检报告中的血脂四项检验项目及结果临床描述与国际规范术语集观测指标标识符逻辑命名与编码系统(Logical Observation Identifiers Names and Codes,LOINC)、医学系统命名法—临床术语(Systematized Nomenclature of Medicine-Clinical Terms,SNOMED CT)英文术语集的对齐关系。庞綑等[4]选择文本相似度算法WMD实现康复量表与国际功能、残疾和健康分类(International Classification of Functioning, Disability and Health,ICF)编码之间的对齐工作。尹帅龙等[5]提出了一种利用Skip-gram词向量模型,以余弦相似度作为输出,通过分析语义信息完成口语化疾病名称与国际规范疾病术语集ICD-11专业术语的映射。

在国际医学术语标准化工具的发展逐渐成熟和稳定后,多语种和多标准术语集的交叉映射和集成融合应运而生。例如,LOINC致力于完成与SNOMED CT的对齐编码,同样SNOMED CT也实现了与ICD-9-CM、ICD-10、ICD-10-CM、ICF等的交叉映射。随着计算机科学技术的迅猛发展,NGUYEN等[6]研究了一种采用自然语言处理(Natural Language Processing,NLP)方法指导SNOMED CT和ICD-10-AM(澳大利亚修改)之间的编码映射,通过评估和分析,基于NLP的计算机辅助编码(Computer Assisted Coding,CAC)相较于其他的机器学习方法取得了较好的结果。DRENKHAHN等[7]基于LOINC和SNOMED CT的本体论工具对实验室数据进行聚合和可视化。

国际标准术语集的更新维护频繁,其手工映射非常耗时,虽然已有开发的映射工具用于加速各国际标准术语的对齐过程,但是当需要映射大规模术语时仍可行性不高,对术语对齐任务有一定的挑战。本文利用自然语言的语义信息,提出一种基于语义相似度的医学术语集编码对齐方法,运用预训练语言模型PubMedBERT对相关的术语描述和概念进行编码对齐,期望为医学领域的对齐工作提供方法参考,减少数据协调过程中的人工参与度,辅助专业人员的编码决策。

2 技术介绍(Technology introduction)

2.1 预训练模型

PubMedBERT是由微软研究人员提出的一种针对生物医学NLP的领域特定语言模型[8]。BERT预训练语言模型基于双向Transformer编码器,以未标注维基百科或其他通用大规模语料进行训练,缺乏专业医学方面的知识体系,在处理生物医学领域的自然语言任务时效果不尽如人意。BERT是混合领域的预训练模型,而PubMedBERT是一种针对特定领域预训练的新范式。PubMedBERT在BERT结构的基础上直接以PubMed生物医学语料摘要和PubMedCentral全文文章进行训练,此外在PubMedBERT中也具有一些特定的技术,例如过滤掉非医疗领域的语料库,加入更多的医学术语以提高PubMedBERT模型的性能,故在生物医学领域更受青睐[9]。

PubMedBERT的基本架构包括四个部分:输入层、编码器层、预测层和输出层。在输入嵌入层中,神经网络会将输入文本表示为一维词向量,包括标记嵌入、段落嵌入和位置嵌入。编码器层中使用了多层的Transformer编码器,每个编码器包括自注意力层和全连接前馈层,用于对输入序列进行特征提取和表示学习。这些编码器能够捕获输入序列中的上下文关系,将其编码为固定长度的向量表示。预测任务包括MLM(Masked Language Modeling)机制和NSP(Next Sentence Prediction)机制[10]。通过MLM随机屏蔽输入序列中的一些Token,利用未遮蔽词预测出被屏蔽的Token内容。输出层是输入序列的向量表示,该向量融合了全文的语义信息。PubMedBERT模型架构图如图1所示。

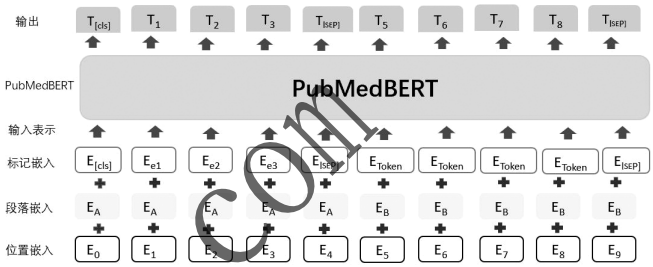

图1 PubMedBERT模型架构图

Fig. 1 Model architecture of PubMedBERT

2.2 传统文本向量化

基于神经网络的计算模型和传统的计算模型是处理自然语言语义相似度问题时的两大主流方向。传统的文本相似度计算中的代表性算法有TF-IDF模型、LSI/LSA模型、BM25模型、VSM模型等,它们的共同特性是利用传统的统计词频和相似度计算公式实现语义相似度计算,而非借助神经网络。例如,BM25算法通过对查询句子进行语素解析,获取Query中的分词 q_i,对于搜索到的文档 d,计算Query中每个分词与 d 的相关性,将所得分数进行加权求和,最终计算得到Query与检索文档 d 的相关性分数[11]。BM25算法如公式(1)所示,W_i 表示第 i 个词的权重,即 IDF,如公式(2)所示,N 表示索引中全部文档数,df_i 表示包含 q_i 的文档数量。

$$Score(Q,d)=\sum_{i=1}^{n}W_i\cdot R(q_i,d) \tag{1}$$

$$IDF(q_i)=\log\frac{N-df_i+0.5}{df_i+0.5} \tag{2}$$

非监督学习算法潜在语义分析(Latent Semantic Analysis,LSA)在信息检索领域的应用广泛,通过LSA得到单词-文档矩阵后,利用奇异值分解(SVD)对矩阵降维去噪[12]。VSM(向量空间模型)利用空间中的特征向量度量文本内容,向量中的每个元素表示在整个集合中出现词项的频率。

2.3 相似度计算

余弦相似度(Cosine Similarity)是判断两个文本之间相似度的一种便捷、有效的方法。通过计算两个向量夹角的余弦值来衡量两个向量间差异的大小,余弦值越接近1,就表明两个向量越相似。对于空间中的任意 n 维向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$

和 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)$,余弦相似度的计算如公式(3)所示:

$$\cos(\boldsymbol{x},\boldsymbol{y})=\frac{\sum_{k=1}^{n}x_k y_k}{\sqrt{\sum_{k=1}^{n}x_k^2}\sqrt{\sum_{k=1}^{n}y_k^2}} \tag{3}$$

3　实验及结果分析(Experiment and result analysis)

3.1　技术路线

首先对待对齐的医学术语的描述文本分别进行预处理操作,对文本进行缩略词扩展,丰富短文本的内容,筛选剔除不必要的特殊符号并进行单词归一化处理,达到降噪的目的;其次运用预训练语言模型进行术语文本相似度计算;最后实现术语匹配对齐。本文以国际上应用范围较广的医学术语体系 LOINC 和 SNOMED CT 作为测试语料,将 SNOMED CT 中的术语文本作为被匹配对象,LOINC Part 中的短文本作为目标术语,选择与被匹配对象最高的前 K 个 LOINC 目标的余弦相似度值,以 LOINC 官方提供的映射结果作为评价标准,TOP-K 精度作为性能度量。医学术语对齐流程图如图 2 所示。

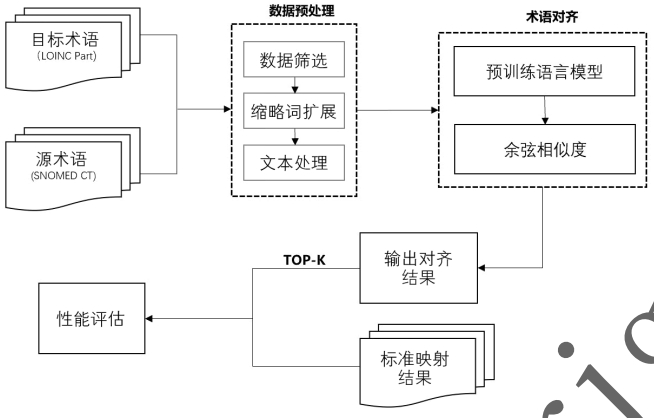

图 2　医学术语对齐流程图

Fig. 2 Medical terminology alignment flowchart

3.2　实验数据

目标术语选自 LOINC 文本数据,使用 2022 年 8 月由美国雷根斯基夫研究院所发布的版本 LOINC 2.73,该版本共包含 99 079 个 LOINC 术语,Part 文件中唯一的 LP-Code 共 69 880 条。源术语为 SNOMED CT,是由国际医疗术语标准开发组织发布的版本 20220301,标准映射文件选自 LOINC 2.73 中的 Mapping 文件,缩略词表下载自 2022 年 7 月份更新的 LOINC 缩写词和首字母缩略词的字母顺序列表,共 904 条,此表可在 LOINC 官方网站获取,同时扩充医学缩略词表,使其能够涵盖更多的缩略词信息。

3.3　数据预处理

(1)数据筛选。在匹配数据中,因 SNOMED CT 的数据量庞大,若全部用于对齐工作,会给模型操作带来很大的负担,因此只抽取部分首选术语作为映射文本。由于标准映射文件包含 LOINC Part 和外部编码系统中的概念之间的映射,如 SNOMED CT 和 RadLex,因此剔除不必要的 RadLex 部分,仅保留与 SNOMED CT 等效的 5 918 条映射结果。

(2)缩略词扩展。医学术语普遍具有相似的特点,它们都具有明确的定义和标准化,医学术语的缩写大都遵循领域内的标准原则。缩略词表主要包含生物医学领域中的各种缩写词,如缩略词"PPP"可以扩展为"Platelet Poor Plasma","RBCCo"可以扩展为"Red Blood Cells Cord","VRatCnt"可以扩展为"Volume Rate Content",故将源文本和目标文本中含有缩略词的部分进行扩展后,再进行文本相似度计算。此方法可以更好地理解语义,提高相似度匹配的准确率。

(3)文本处理。特殊符号在自然语言中是普遍存在的,但它们可能不会为文本含义增加太多价值甚至会干扰文本语义的理解。医学领域的描述性文本中通常存在医学标识符,这些符号可能会对术语对齐的结果产生一定的干扰和迷惑作用。因此,首先对英文文本进行大小写处理,将其全部转换为小写字母,筛选并删除不必要的符号,其次使用自然语言处理工具包(Natural Language Toolkit, NLTK)进行文本分词并删除停用词,最后进行词干提取,使单词归一化。

3.4　实验环境及评价指标

通过 Python 3.10.8 编写实验代码,基于 PyTorch 框架实现,硬件环境为 Intel(R) Xeon(R) W-2245,显卡为 RTX 2080,操作系统为 Ubuntu 18.04.6 LTS,运用准确率(Accuracy,A)作为性能评估的方法,它被定义为正确目标在 TOP-K 模型预测中的样本的百分比,如公式(4)所示:

$$A=\frac{n_k}{N} \tag{4}$$

3.5　结果和讨论

本文设计了三个对比实验。医学文本数据中普遍包含有大量的符号和缩略词,为探究缩略词对术语对齐是否有干扰作用,设计实验一。传统的文本相似度计算模型主要是利用统计词频和相似度计算公式实现相似度计算,不需要借助神经网络,而深度学习方法的非线性建模能力更强,可以更好地利用语义信息,因此设计实验二对比传统模型与预训练模型对准确率的影响。不同的预训练模型的预训练方式和训练时所用的语料有所不同,为探究不同预训练模型对性能结果的影响,设计实验三,以期找出适应此文本对齐的效果最好的模型。

3.5.1　缩略词扩展前后对比实验

由图 3 可以得出,缩略词扩展后,PubMedBERT 模型 TOP1、TOP3、TOP5、TOP10 的准确率均显著提升,由实验结果分析可见,缩略词扩展能明显地提升语义的表达能力,对文本匹配具有较大的贡献。

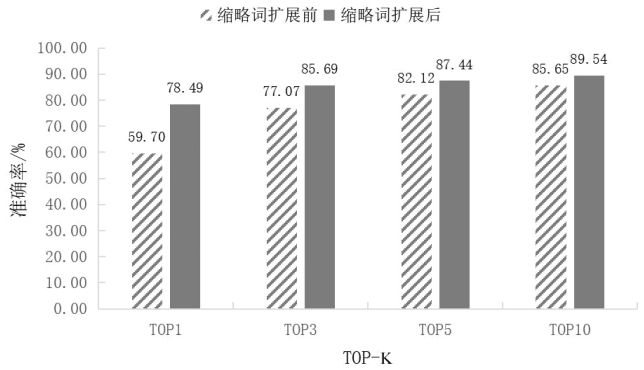

图 3　缩略词扩展前后对比图

Fig. 3 Comparison chart before and after abbreviation expansion

3.5.2　传统模型与深度学习模型性能比较

本文对比分析了缩略词扩展后传统模型 BM25、LSA、VSM 和 PubMedBERT 的性能表现,由表 1 可以看出,传统模

型的性能表现较预训练语言模型要差很多，这可能是由于传统模型不能较好地理解单词之间的语义关系。同时，能证明预训练语言模型更符合术语对齐推荐编码的需求。

表1 各模型的性能比较

Tab.1 Performance comparison of models

模型	准确率/%			
	TOP1	TOP3	TOP5	TOP10
LSA	46.40	48.90	50.10	51.48
VSM	54.91	62.26	65.20	66.00
BM25	54.93	67.71	71.10	74.83
PubMedBERT	78.49	85.69	87.44	89.54

3.5.3 不同预训练语言模型对比实验

分析表2可知，当进行缩略词扩展之后，Bio_ClinicalBERT模型与PubMedBERT模型TOP3、TOP5、TOP10的准确率相差并不大，但在所有模型中，PubMedBERT模型TOP1的准确率最高。不同模型表现出的性能的区别与它们本身训练所使用的语料库有直接的关联。本实验所使用的BERT模型、Bio_ClinicalBERT模型、PubMedBERT模型训练使用的语料库分别为通用语料库、临床文本和MIMIC记录、生物医学语料，它们在词汇分布等方面均有不同之处，造成模型的具体性能表现有所差异[13]。从总体性能表现来看，PubMedBERT模型在术语对齐任务中的表现最优异。

表2 各深度学习模型的实验结果

Tab.2 Experimental results of each deep learning model

模型	准确率/%			
	TOP1	TOP3	TOP5	TOP10
BERT	67.61	80.69	83.25	85.92
Bio_ClinicalBERT	73.99	84.30	86.68	88.67
PubMedBERT	78.49	85.69	87.44	89.54

4 结论(Conclusion)

针对医疗领域中医学术语标准化程度低、缺乏语义互操作性的现象，本文提出了一种运用生物医学领域的预训练语言模型PubMedBERT和余弦相似度计算，结合医学缩略词扩展、符号筛选等预处理方式，进行术语对齐TOP-K准确率的测评。实验结果证明，缩略词扩展对医学术语的语义理解有显著作用，基于深度学习的预训练模型的术语对齐结果优于传统的相似度计算模型，并且缩略词扩展后，PubMedBERT模型相较于BERT模型及其变体取得了更好的效果，表明此方法用于文本对齐和推荐编码的巨大潜力。将PubMedBERT模型与缩略词扩展方法相结合，有利于专业人员的学术探讨并能为编码员提供一定的帮助，同时为医学文本对齐任务的完成提供了一种新思路。下一步的工作应注重增强模型TOP1的准确率，更好地提升映射的质量。同时，应考虑该方法在不同数据集上的效果，提高模型的泛化能力，使其能够适用于更多类似的医学术语之间的对齐任务。

参考文献(References)

[1] ROSSANDER A，LINDSKÖLD L，RANERUP A，et al. A state-of-the art review of SNOMED CT terminology binding and recommendations for practice and research[J]. Methods of Information in Medicine，2021，60(S2)：e76-e88.

[2] 任慧玲，李晓瑛，邓盼盼，等. 国际医学术语体系进展及特色优势分析[J]. 中国科技术语，2021，23(3)：18-25.

[3] 赵云松，杨鹏，张林，等. 血脂四项检验项目及结果临床描述与国际规范术语集映射[J]. 中国卫生信息管理杂志，2017，14(6)：862-867.

[4] 厐細，郑建立. 基于文本相似度的康复量表ICF映射研究[J]. 软件导刊，2022，21(4)：181-185.

[5] 尹帅龙，夏晨曦. 口语化疾病名称向国际规范疾病术语集的映射研究[J]. 中华医学图书情报杂志，2020，29(1)：22-27.

[6] NGUYEN A N，TRURAN D，KEMP M，et al. Computer-assisted diagnostic coding：effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings[J]. AMIA Annual Symposium Proceedings/AMIA Symposium，2018：807-816.

[7] DRENKHAHN C，DUHM-HARBECK P，INGENERF J. Aggregation and visualization of laboratory data by using ontological tools based on LOINC and SNOMED CT[C]// Studies in Health Technology and Informatics. Proceedings of the 17th World Congress on Medical and Health Informatics. Netherlands：IOS Press BV，2019：108-112.

[8] 董森，苏中琪，周晓北，等. 利用Text-CNN改进PubMedBERT在化学诱导性疾病实体关系分类效果的尝试[J]. 数据分析与知识发现，2021，5(11)：145-152.

[9] GU Y，TINN R，CHENG H，et al. Domain-specific language model pretraining for biomedical natural language processing[J]. ACM Transactions on Computing for Healthcare，2022，3(1)：1-23.

[10] SIGIRCI I O，BILGIN G. Spectral-spatial classification of hyperspectral images using BERT-based methods with Hyper-SLIC segment embeddings[J]. IEEE Access，2022，10：79152-79164.

[11] 李楠，陶宏才. 一种新的融合BM25与文本特征的新闻摘要算法[J]. 成都信息工程大学学报，2018，33(2)：113-118.

[12] 郝秀慧，方贤进，杨高明. 基于TFIDF+LSA算法的新闻文本聚类与可视化[J]. 计算机技术与发展，2022，32(7)：34-38，45.

[13] TUNG N Y，HU H W，CHANG T W，et al. Multi-model comparison for classification of medical records using the BioBERT models[C]//Institute of Electrical and Electronics Engineers. Proceedings of the 2022 IEEE 4th Eurasia Conference on Biomedical Engineering，Healthcare and Sustainability(ECBIOS). Piscataway：IEEE Computer Society，2022：221-224.

作者简介：

王怡茹(1998-)，女，硕士生。研究领域：自然语言处理。

郑建立(1965-)，男，博士，副教授。研究领域：医学信息系统与集成技术。本文通信作者。

周浩然(1996-)，男，硕士生。研究领域：自然语言处理。