文章编号：2095-2163(2023)06-0184-05 中图分类号：TP391.1；R197.32 文献标志码：A

基于改进 MC-Bert 的 ICD 编码映射方法研究

周浩然，郑建立
（上海理工大学 健康科学与工程学院，上海 200093）

摘 要：当前国内医疗体系中存在多种版本的 ICD 编码，编码映射是保障编码数据跨版本准确性的主要手段。本文提出了改进 MC-Bert 的 ICD 编码映射方法，基于医学名称与编码组合的语义相似度计算，在 ICD-10 国标 2020 版和医保版 2.0 对照库的非重名数据中 top1、top3、top5 3 种匹配精度的准确率分别达到 89.6%、96.6%、97.6%，在 ICD-9 团标 2020 版和医保版 2.0 对照库的非重名数据中 top1、top3、top5 3 种匹配精度的准确率分别达到 91.5%、96.8%、97.8%，为加速推进医疗数据标准化提供技术基础。

关键词：ICD 编码；语义相似度；编码映射

Research on ICD code mapping method based on improved MC-Bert

ZHOU Haoran, ZHENG Jianli
(University of Shanghai for Science and Technology, School of Health Science and Engineering, Shanghai 200093, China)

【Abstract】 Currently multiple versions of ICD codes exist in the domestic medical system, and code mapping is the main means to ensure the accuracy of coded data across versions. This paper proposes an improved MC-Bert ICD code mapping method based on the semantic similarity calculation of medical names and code combinations. In the non-duplicate data from comparison databases of the ICD-10 national standard 2020 version and the medical insurance version 2.0, the top1、top3、top5 accuracy of our method reach 89.6%、96.6% and 97.6%, respectively. In the non-duplicate data from comparison databases of the ICD-9 group standard 2020 version and the medical insurance version 2.0, the top1、top3、top5 accuracy of our method reach 91.5%、96.8%, and 97.8%.

【Key words】 ICD code; semantic similarity; code mapping

0 引 言

疾病和相关健康问题的国际统计分类（International Statistical Classification of Diseases and Related Health Problems, ICD）由世界卫生组织创立，用来确定全球卫生趋势和统计数据的一种医疗编码体系国际标准。该体系由表 1 所示的医学编码及对应医学名称组成最小描述单元，涉及到手术、疾病、诊断等医疗环节，对生物医学领域如医学知识实体对齐、医疗标准化、临床路径等研究起着重要作用，同时也作用于医保结算、医疗监督等领域。

当前，国内医疗体系中存在着多种本地化的 ICD 编码版本，且大部分基于 ICD-9 和 ICD-10。虽然部分机构发布了某版本与另一版本的映射，但不论是从映射版本的数量以及更新速度都不尽如人意。除此以外，各个医疗机构还存在各自定义的院内码，这更对医疗数据的一致性提出了挑战。

表 1 ICD 编码示例

Tab. 1 Examples of ICD code

编码名称	编码
古典生物型霍乱	A00.000x001
孤立性蛋白尿伴肾小球损害	N06.900
遗传性肾病伴有轻微的肾小球异常，不可归类在他处者	N07.000
心脏起搏器电极功能异常	T82.102
心脏起搏器电极移位	T82.103

目前，医学编码相关的研究大多集中在病案的命名实体识别和编码领域，如夏等[1]基于深度学习技术实现电子病历的实体识别；庞等[2]基于文本相似度实现了康复量表与 ICF（International Classification of Functioning, Disability and Health）编码的映射。此外，专业医生也就各自专业领域 ICD

作者简介：周浩然（1996-），男，硕士研究生，主要研究方向：医学信息系统与集成技术；郑建立（1965-），男，博士，副教授，主要研究方向：医学信息系统与集成技术。
通讯作者：郑建立 Email：zhengjianli163@163.com

收稿日期：2022-06-30

编码的合理性进行了讨论，如叶[3]等对 ICD-10 在眼挫伤的分类编码讨论；许等[4]对 ICD-10 编码在癫痫方面的质量分析。

实现 ICD 映射的方式往往需要大量的人工分级、字典映射等传统方式，而基于语义相似度的方法较少。随着蕴含大量生物医学领域先验知识的预训练模型 MC-Bert（Meta-Controller BERT）的出现，中文医学文本可以转化为更加稠密和准确的向量表示，在此基础上本文提出一种基于改进 MC-Bert 的 ICD 编码映射方法，该方法通过语义相似度在现有的 ICD 版本映射库中进行匹配实验，在不同匹配精度下的准确率均达到较高水平。

1　改进的 MC-Bert 模型

改进的 MC-Bert 是一种利用白化处理优化 MC-Bert 编码输出的无监督学习模型，其结构图如图 1 所示。

首先，由于 ICD 中的名称部分既有较短小的词语如霍乱，也有较长的句子如“遗传性肾病伴有轻微的肾小球异常，不可归类在他处者”，本文将其统一填充为相同长度的句子，输入 MC-Bert 进行编码；其次，将两个文档中编码名称的输出矩阵拼接，作为白化处理的输入，计算获得消除各向异性后的句向量；最后，将两文档的句向量两两计算余弦相似度，依次进行排序获得 Top5，输出用于验证。

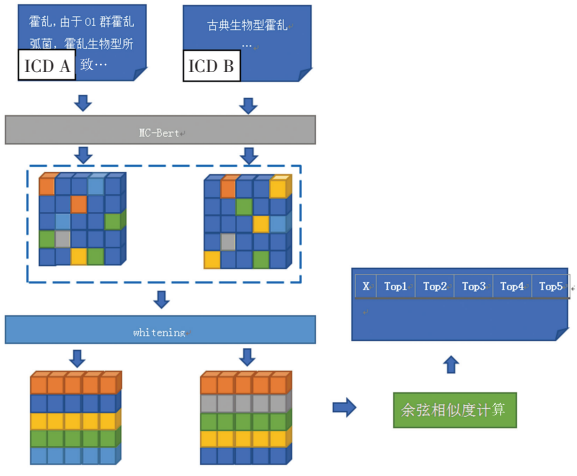


图 1　改进的 MC-Bert 模型结构图

Fig. 1　Model structure diagram of improved MC-Bert

1.1　预训练语言模型 MC-Bert

MC-Bert 由 Zhang 等[5]提出，训练过程如图 2 所示。以 BERT 作为基础模型，使用大量生物医学领域语料进行训练，包含许多生物医学领域先验知识。虽然预训练语言模型在各项语言任务中性能均有大幅的提升，但 Gao 等[6]发现，其在词向量方面仍存在各向异性，导致模型出现语义表达的退化问题。

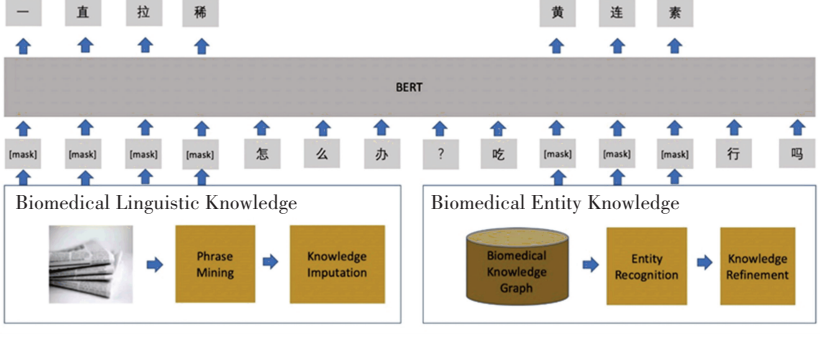


图 2　MC-Bert 的训练过程

Fig. 2　The training process of MC-Bert

1.2　白化处理

白化处理是一种预处理方法，由 Su 等[7]首先引入以解决预训练模型语义表达的退化问题，其具体操作是将文档中 N 条句子经过预训练模型的编码层输出为向量集合 $\{x_i\}_{i=1}^{N}$，然后将此集合经过如式(1)的线性变换，转变为均值为 0 且协方差矩阵为单位矩阵的向量集合 $\{\tilde{x}_i\}_{i=1}^{N}$。

$$\tilde{x}_i = (x_i - \mu)\,\omega \tag{1}$$

其中，μ 代表平移系数，ω 代表缩放系数。

为了实现 $\tilde{x}_i$ 的均值为 0，则 μ 需要满足式(2)：

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

而 $\{x_i\}_{i=1}^{N}$ 的协方差矩阵 $\boldsymbol{\Sigma}$ 满足式(3)：

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^{\mathrm{T}}(x_i - \mu) \tag{3}$$

转换后 $\{\tilde{x}_i\}_{i=1}^{N}$ 的协方差矩阵 $\tilde{\boldsymbol{\Sigma}}$ 与 $\boldsymbol{\Sigma}$ 的关系为式(4)：

$$\tilde{\boldsymbol{\Sigma}} = \omega^{\mathrm{T}}\boldsymbol{\Sigma}\omega \tag{4}$$

由于 $\tilde{\boldsymbol{\Sigma}}$ 为单位矩阵，则式(4)等价于式(5)：

$$\omega^{\mathrm{T}}\Sigma\omega = I \tag{5}$$

由此可得到 Σ 满足式(6):

$$\Sigma = (\omega^{-1})^{\mathrm{T}}\omega^{-1} \tag{6}$$

由于协方差矩阵是正定对称矩阵,因此 Σ 满足式(7)所示的奇异值分解:

$$\Sigma = U\Lambda U^{\mathrm{T}} \tag{7}$$

其中, U 是 $\Sigma\Sigma^{\mathrm{T}}$ 的特征向量矩阵, Λ 为对角矩阵

由式(6)、式(7)联立,可以得到式(8):

$$(\omega^{-1})^{\mathrm{T}}\omega^{-1} = U\Lambda U^{\mathrm{T}} = U\sqrt{\Lambda}\sqrt{\Lambda}U^{\mathrm{T}} = (\sqrt{\Lambda}U^{\mathrm{T}})^{\mathrm{T}}\sqrt{\Lambda}U^{\mathrm{T}} \tag{8}$$

最终可得到 ω 满足式(9):

$$\omega = U\sqrt{\Lambda^{-1}} \tag{9}$$

1.3 余弦相似度

余弦相似度是一种常用的计算文本相似度的方法,计算公式(10):

$$\cos(x,y) = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d} y_i^2}} \tag{10}$$

其中, x,y 代表两条句向量; d 代表句向量的长度; x_i,y_i 代表 x,y 在下标为 i 处的值。

余弦相似度的值越接近 1,两个句子的相似度越高。

2 实验方法和评价指标

2.1 实验数据

本文采用 ICD-10 国标 2020 版和医保版 2.0 对照库以及 ICD-9 团标 2020 版和医保版 2.0 对照库作为实验数据,其实例见表 2。

表 2 实验数据示例

Tab. 2 Examples of experimental data

国标 2020 版编码	国标 2020 版名称	医保 2.0 编码	医保 2.0 名称
A00.000	霍乱,由于 O1 群霍乱弧菌,霍乱生物型所致	A00.000x001	古典生物型霍乱
A00.100	霍乱,由于 O1 群霍乱弧菌,埃尔托生物型所致	A00.100x001	埃尔托生物型霍乱
A09.008	新生儿感染性腹泻	A09.900x005	新生儿腹泻
A16.502	结核性胸腔积液	A16.500x004	结核性胸膜炎
A22.100	肺炭疽	A22.102+J17.0 *	炭疽肺炎

2.2 实验环境及评价指标

改进的 MC-Bert 通过 python 3.9.7,基于 PyTorch 框架实现;硬件环境为 Intel Core i7-11700,显卡为 RTX 3060,显存 12 G,操作系统为 window 10。运用 Top-K 准确率(Accuracy)评估方法性能,计算如公式(11)所示:

$$A = \frac{n_k}{N} \tag{11}$$

其中, n_k 是前 k 个候选项中包含正确项的个数, N 是映射条目的总数。

2.3 实验设计

本文涉及到使用不同版本的 ICD 名称进行相似度计算,但不同版本的 ICD 之间可能存在大量重复的医学名称,会干扰不同医学名称间的相似度匹配结果,因此设计实验(1);ICD 编码数据蕴含丰富的医学知识,注入这类数据或可提高模型匹配的准确率,因此设计实验(2);为了验证改进 MC-Bert 与其他模型在匹配准确率上确有提升,因此设计实验(3)。

(1)重名项对非重名项匹配的干扰评估实验:从 ICD-10 国标 2020 版中筛选出与医保 2.0 版医学名称不重复的 1 773 条数据,分别与去除重名项的、包含重名项的医保 2.0 版数据进行匹配实验。

(2)医学编码注入与否的对比实验:编码部分包含类目、亚目、细目、附加码,分别代表不同范围的医学知识范畴。ICD-9 团标版中筛选出非重名项 1 289 条,医保 2.0 版中非重名项 1 255 条,分为编码不注入、整条编码注入、拆分三类编码分别注入 3 种数据进行对比实验,3 种实验数据示例见表 3。

表 3 三组实验数据示例

Tab. 3 Three sets of experimental data examples

数据类型	数据示例
编码不注入	沙门菌肾小管-间质疾病
整条编码注入	沙门菌肾小管-间质疾病,A02.206+N16.0 *
拆分三类编码注入	沙门菌肾小管-间质疾病,A02,206,A02.206,A02.206+N16.0 *

（3）改进 MC-Bert 与其他模型的对比实验：在数据去重和拆分三类编码注入后，在 ICD-10 国标 2020 版和医保 2.0 版对照库以及 ICD-9 团标版和医保 2.0 版对照库中，就改进 MC-Bert 和 TF-IDF（Term Frequency-Inverse Document Frequency）、LSI（Latent Semantic Indexing）、MC-Bert、VSM（Vector Space Model）模型的表现进行对比。

3　结果分析和总结

3.1　重名项对非重名项的扰动评估实验

ICD-10 国标 2020 版与医保 2.0 版重名项对非重名项匹配的干扰结果见表 4。由此实验证明，重名项对非重名项匹配的干扰影响较大，因此需要将不同版本 ICD 中的重名项和非重名分开匹配。同时，也验证了改进 MC-Bert 在非重名项之间依旧保有较高的准确率。

表 4　重名项对非重名项匹配的干扰结果

Tab. 4　The experimental results of the perturbation evaluation of the duplicated items to the non-duplicated items　%

模型	top1 准确率	top3 准确率	top5 准确率
改进 MC-Bert$_{含重名项}$	56.40	69.48	74.96
改进 MC-Bert$_{去重名项}$	81.33	90.58	93.17

3.2　医学编码注入与否的对比实验

CD-9 团标版中非重名项 1 289 条，医保 2.0 版中非重名项 1 255 条分别对文本中的英文、符号进行预处理后，分为编码不注入、整条编码注入、拆分三类编码注入的性能对比实验结果见表 5。由此实验证明，拆分编码为类目、亚目、细目三级注入医学名称中可显著提升准确率，因而结合医学名称与三级编码是最为合理的语义匹配方案。

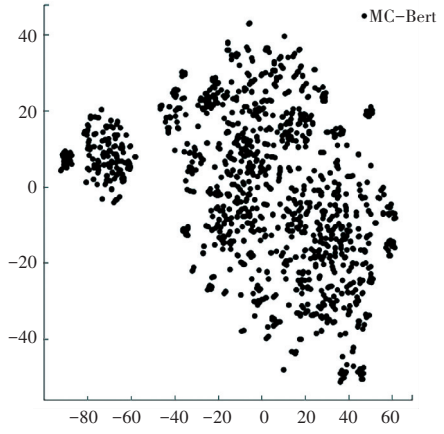
表 5　三种实验数据的性能对比结果

Tab. 5　Comparison of experimental results of three groups of experimental data　%

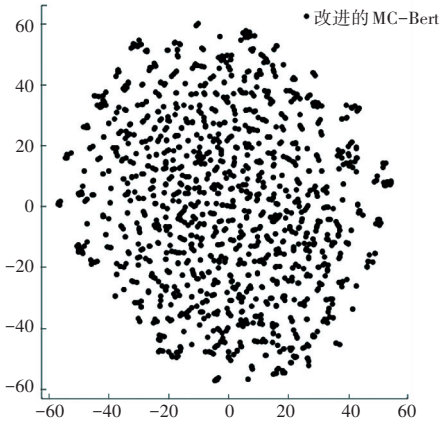
模型	top1 准确率	top3 准确率	top5 准确率
改进 MC-Bert$_{编码不注入}$	78.95	89.20	91.92
改进 MC-Bert$_{整条编码注入}$	86.02	94.72	96.19
改进 MC-Bert$_{拆分三类编码注入}$	91.53	96.81	97.82

3.3　改进 MC-Bert 与其他模型的对比实验

通过将 ICD-9 团标版中非重名的 1 289 条向量进行 t-SNE 降维，对降维后的向量进行可视化，得到如图 3 所示的向量分布对比图，可见改进 MC-Bert 相较 MC-Bert 能够有效的将重叠的向量分散开来，拥有更好的语义表达能力，提升语义相似度检索的敏感度。



（a）MC-Bert



（b）改进 MC-Bert

图 3　向量分布对比图

Fig. 3　Comparison of vector distributions

同时本文也对加进 MC-Bert 与其他模型在 ICD-10 国标 2020 版和医保版 2.0 对照库以及 ICD-9团标 2020 版和医保版 2.0 对照库上非重名项的准确率进行比较，结果见表 6、表 7。

表 6　ICD-9 团标 2020 版和医保版 2.0 映射的对比实验结果

Tab. 6　Comparison experiments of ICD-9 group standard 2020 version and medical insurance version 2.0 mapping　%

模型	top1 准确率	top3 准确率	top5 准确率
TF-IDF$_{拆分三类编码注入}$	81.59	84.86	86.25
LSI$_{拆分三类编码注入}$	85.71	90.76	92.39
MC-Bert$_{拆分三类编码注入}$	87.18	92.93	94.33
VSM$_{拆分三类编码注入}$	93.32	95.57	95.88
改进 MC-Bert$_{拆分三类编码注入}$	91.53	96.81	97.82

表 7 ICD-10 国标 2020 版和医保版 2.0 映射的对比实验结果

Tab. 7 Comparison experiments of ICD-10 national standard 2020 version and medical insurance version 2.0 mapping %

模型	top1 准确率	top3 准确率	top5 准确率
$TF-IDF_{拆分三类编码注入}$	66.94	73.49	75.23
$LSI_{拆分三类编码注入}$	79.63	86.74	89.17
$MC-Bert_{拆分三类编码注入}$	87.98	94.58	95.93
$VSM_{拆分三类编码注入}$	91.48	93.68	94.24
改进 $MC-Bert_{拆分三类编码注入}$	89.62	96.61	97.57

可以看到改进后的 MC-Bert 模型与其他模型相比,除了在 top1 匹配精度下的准确率方面低于 VSM 模型外,其他匹配精度下的准确率较其他模型有较大提升。

4 结束语

本文提出来一种基于改进 MC-Bert 的 ICD 编码映射方法,通过实验证明了该方法相较其他模型在准确率方面有较大的提升,为医学编码领域的智能化映射提供了一种思路。

参考文献

[1] 夏宇彬, 郑建立, 赵逸凡, 等. 基于深度学习的电子病历命名实体识别 [J]. 电子科技, 2018, 31(11): 31-4,7.
[2] 庞絅, 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.
[3] 叶文琦, 方浩, 林明色. 眼挫伤的 ICD-10 分类编码探讨 [J]. 中国病案, 2022, 23(3): 25-27.
[4] 许莹, 杨静怡, 彭蓉, 等. 癫痫 ICD-10 编码质量分析 [J]. 中国病案, 2022, 23(2): 54-57.
[5] ZHANG N, JIA Q, YIN K, et al. Conceptualized representation learning for chinese biomedical text mining [J]. ArXiv preprint arXiv 2008.10813,2020.
[6] GAO J, HE D, TAN X, et al. Representation degeneration problem in training natural language generation models [J]. ArXiv preprint arXiv 1907.12009,2019.
[7] SU J, CAO J, LIU W, et al. Whitening sentence representations for better semantics and faster retrieval [J]. ArXiv preprint arXiv 2103.15316,2021.

(上接第 183 页)

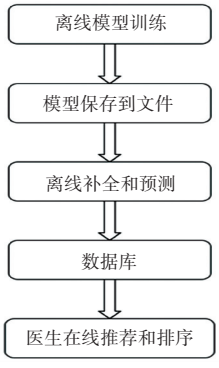


图 3 信息推荐系统结构

Fig. 3 Information recommendation system structure

4 结束语

本文提出的基于标签预测和奇异值分解的医生推荐系统,能够实现患者在就医时寻找所需医生的目的,解决就医时选择合适医生的问题,对于患者和医生都起到非常重要的作用。医生推荐系统是一项持续改进的工程,未来还需要在算法和排序等方面进行更加深入的改进,使其在现实就医过程中发挥出更大的作用和效果。

参考文献

[1] 高山,刘炜,崔勇,等. 一种融合多种用户行为的协同过滤推荐算法[J]. 计算机科学,2016,43(9):227-231.
[2] 周岩,雷世尧,张灿. 面向移动健康医疗系统的多层二分网络推荐算法[J]. 中国科学院大学学报,2017,34:112-118.
[3] 黄山山. 协同过滤推荐算法的关键性问题研究[D]. 山东:山东大学,2016.
[4] 曹一鸣. 协同过滤推荐瓶颈问题综述[J]. 软件,2012,33(12):315-321.
[5] 杨文显. 项目评审专家协同推荐方法的研究及应用[D]. 杭州:杭州电子科技大学,2016.