临床诊断编码技术评测数据集及基线模型概述*

罗庚欣 康 波

（哈尔滨工业大学（深圳） 深圳 518055） （医渡云（北京）技术有限公司 北京 100191）

彭 浩 熊 英 汤步洲

（哈尔滨工业大学（深圳） 深圳 518055） （哈尔滨工业大学（深圳）鹏城实验室 深圳 518055）

〔摘要〕 介绍第八届中国健康信息处理会议（CHIP 2022）发布的临床诊断编码评测任务、临床诊断编码技术评测数据集，详细阐述评测所使用数据集来源和组成、基线模型构建思路以及实验结果，为相关研究提供参考。

〔关键词〕 临床诊断编码；人工智能；自然语言处理；评测数据集

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.12.003

An Overview of Clinical Diagnosis Coding Technique Evaluation Dataset and Basline Model *LUO Gengxin, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China; KANG Bo, Yidu Cloud (Beijing) Technology Co. Ltd., Bejing 100191, China; PENG Hao, XIONG Ying, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China; TANG Buzhou, Peng Cheng Laboratory, Harbin Institute of Technology, Shenzhen 518055, China*

〔**Abstract**〕 The paper introduces the clinical diagnosis coding evaluation task and clinical diagnosis coding technique evaluation dataset released at the 8th China Health Information Processing Conference (CHIP 2022), and elaborates the source and composition of the dataset used in the evaluation, the construction of baseline model and experimental results, so as to provide references for related research.

〔**Keywords**〕 clinical diagnosis coding; Artificial Intelligence (AI); Natural Language Processing (NLP); evaluation dataset

1 引言

在医疗健康领域，大量医疗文本数据以电子病历的形式存储。这些电子病历中包含患者的各类就诊信息，具有重要的统计价值。但电子病历往往是在对患者进行诊断时由医疗工作者输入的，这就导致存在不同病历表达方式不同以及质量不齐等问

〔修回日期〕 2022－11－16

〔作者简介〕 罗庚欣，硕士研究生；通信作者：汤步洲，博士，副教授，博士生导师。

〔基金项目〕 科技创新2030——“新一代人工智能”重大项目“儿科疾病复杂循证知识图谱构建研究”（项目编号：2021ZD0113402）；国家自然科学基金“基于电子病历深层语义分析的临床辅助决策方法研究”（项目编号：62276082）；国家自然科学基金“基于电子病历分析的慢病趋势预测方法研究”（项目编号：61876052）。

题，为后续处理工作带来困难[1]。

临床诊断编码技术能够根据病历样本中与医学相关的概念，为其自动分配一个或多个诊断标准词，这些诊断标准词的产生将有效提高电子病历在后续完成统计等任务时的处理效率。随着自然语言处理技术的发展，越来越多的文本处理任务得到有效解决，其中包括对医疗文本的处理等[2]。在此背景下，如何借助自然语言处理技术推动临床诊断编码技术发展是近期相关研究的热门话题。

自 2018 年开始，中国健康信息处理会议（China Health Information Processing Conference，CHIP）发布了多项医疗健康信息处理相关任务[3-4]，推动了医疗信息与人工智能交叉领域相关研究的发展。2021 年 CHIP 会议发布了临床术语标准化评测任务，该任务要求对中文电子病历中给定的手术原词进行标准化，并映射到标准词表中对应的手术标准词[3]。本文提出的临床诊断编码任务是在临床术语标准化任务基础上做出的进一步拓展和延续。两个任务的区别体现在临床诊断编码任务没有在数据集中对需要标准化的医学术语进行标注，只提供了病历样本中的各类就诊信息，根据就诊信息直接生成相关医学诊断标准词。就诊信息中没有对医学相关术语进行标注，同时某些诊断标准词是根据就诊信息整体的语义而产生，例如标准词“恶性肿瘤靶向治疗”是在“入院诊断”中出现“癌”或“肿瘤”等词语且“药品名称”中出现某种靶向药物的名称时才产生的；而标准词“恶性肿瘤放射治疗”则更倾向于当“医嘱”中出现“放疗”等术语时产生，这就为相关任务带来一定挑战。本文将围绕临床诊断编码任务描述、临床诊断编码技术评测数据集介绍、临床诊断编码基线模型以及实验结果进行阐述。

2 任务描述

2.1 任务背景

2.1.1 *疾病分类与手术操作分类编码* 是对患者疾病诊断和治疗信息的加工过程，是病案信息管理的重要环节[4]。如今病案编码已成为医院科学化、信息化管理的重要依据之一，其在评估医疗质量与医疗效率、设计临床路径方案、医院评审、疾病诊断分级、合理用药监测等方面的应用越来越广泛和深入。

2.1.2 *国际疾病分类* 在诸多的分类方案中，世界上最有影响力且最为普及的是国际疾病分类（International Classification of Diseases，ICD）。ICD 是世界卫生组织（World Health Organization，WHO）制定的国际统一疾病分类方法[5]，是目前国际上通用的疾病分类方法。我国也推出了《疾病分类与代码国家临床版 2.0》和《手术操作分类代码国家临床版 2.0》，并在部分医院中得到应用。

2.2 临床诊断编码评测任务

临床诊断编码评测任务的主要目标是针对中文电子病历进行诊断编码。任务的具体要求如下：给定一次就诊的相关信息，包括入院诊断、术前诊断、术后诊断、出院诊断，以及手术名称、药品名称和医嘱，要求给出本条电子病历所对应的一个或多个诊断标准词，并以《疾病分类与代码国家临床版 2.0》[5]词表作为标准。

3 评测数据

3.1 数据来源

临床诊断编码评测数据集由医渡云（北京）技术有限公司提供，数据由内部专业医学团队基于职业经验进行编写和标注，不包含个人身份信息及基因等遗传资源数据，只使用治疗过程中的诊断等医嘱信息，不是真实场景下的数据，符合隐私保护和伦理要求，该数据集仅限 CHIP 2022 比赛使用。

3.2 数据描述

评测数据主要由两部分组成。数据集 1 共包含 2 700 条样本，数据集 2 共包含 337 条样本。每条样本描述了一次就诊信息，包括入院诊断、出院诊断、手术名称、术前诊断、术后诊断、药品名称、医嘱以及该样本对应的诊断标准词，每种就诊信息均由机器和专业医学团队人工构造得到，见表 1。

表 1 数据集样例

就诊信息（或诊断标准词）	文本
出院诊断	鼻咽癌放化疗后，肝癌综合治疗后，ct3n0m0 iiia/cnlc iib 期，乙型病毒性肝炎病原携带者
手术名称	肝动脉造影术，肝动脉化疗栓塞，肝动脉灌注化疗肝动脉栓塞术，脾动脉栓塞术
术前诊断	肝癌综合治疗后
术后诊断	-
入院诊断	肝癌综合治疗后，ct3n0m0 iiia
药品名称	丁二磺酸腺苷蛋氨酸，仑伐替尼，利多卡因，呋塞米，奥沙利铂，帕洛诺司琼，帕瑞昔布，异丙嗪，恩替卡韦，氟尿嘧啶，氯化钠，氯化钾
医嘱	一支紫管抽血 5 毫升，抗凝管［嘱托］，乙型肝炎 dna 测定（定量），二级护理，今日结帐出院［嘱托］，住院静脉输液，动脉加压注射化疗护理，肝癌介入术后常规护理［嘱托］，肝胆科常规护理［嘱托］
诊断标准词	肝细胞癌；乙肝表面抗原携带者；为肿瘤化学治疗疗程；恶性肿瘤靶向治疗

注：每条样本的诊断标准词可能有多个，各标准词之间以“;”间隔。

此外，数据集 1 中，每 4 条样本的诊断标准词完全相同。数据集 1 中共有 278 个不同标准词，数据集 2 中的诊断标准词均在数据集 1 中出现过。分别针对两个数据集的就诊信息和诊断标准词平均长度进行统计，见表 2。

表 2 数据集中各就诊信息和诊断标准词平均长度

就诊信息（或诊断标准词）	数据集 1 平均字符长度	数据集 2 平均字符长度
出院诊断	24. 84	25. 19
手术名称	11. 61	12. 15
术前诊断	4. 49	4. 48
术后诊断	3. 76	4. 00
入院诊断	21. 62	21. 67
药品名称	89. 69	91. 06
医嘱	917. 21	934. 70
诊断标准词	18. 33	18. 43

由统计结果可见，各就诊信息中，文本最长的是“医嘱”，平均长度超过 900 个字符，且远超其他就诊信息文本长度；其次是“药品名称”，平均长度为 90 个字符，其他就诊信息文本平均长度基本在 25 个字符以内。在本次评测任务中，将数据集 1 作为训练集，将数据集 2 作为测试集。

4 基线模型

4. 1 模型描述

由于该任务与医疗实体标准化任务有相似之处[6]，故采用实体标准化思路构建基线模型。基线模型主要由候选模块和匹配模块两部分构成。其中，候选模块用于为每个样本生成候选诊断标准词，匹配模块将样本就诊信息与候选诊断标准词进行文本匹配，见图 1。

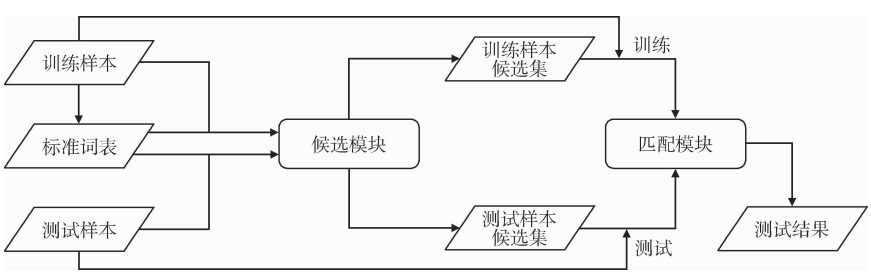

图 1 总体算法流程

4.2 候选模块

在候选模块中，首先统计训练集中所有诊断标准词，并建立标准词词表。在候选过程中，将每条样本中就诊信息，如入院诊断、手术名称、医嘱等合并成为一个文段。利用词频－逆向文件频率（Term Frequency－Inverse Document Frequency，TF－IDF）算法[7]，统计词表中各标准词对每个文段的重要程度并排序，选择重要程度最高的 K 个标准词作为诊断标准词候选集，见图 2。

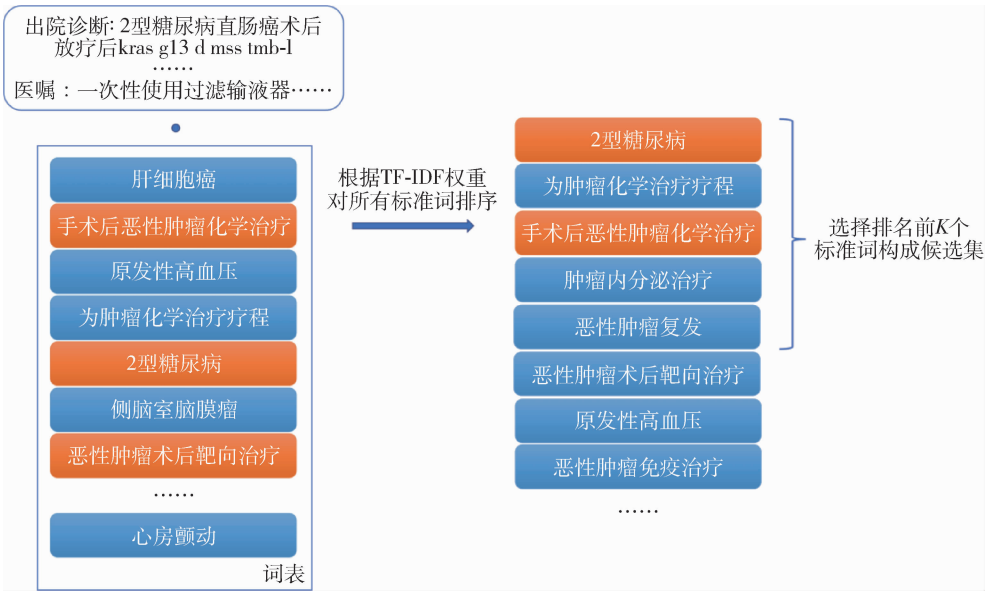

图 2 候选模块算法流程

4.3 匹配模块

匹配模块包含一个预训练语言模型，基线方案采用基于 Transformer 的双向编码器表征（Bidirectional Encoder Representations from Transformers，BERT）[8]作为预训练模型。该模型以每个候选的诊断标准词与对应文段串接后的文本作为输入，输出预测标签，见图 3。

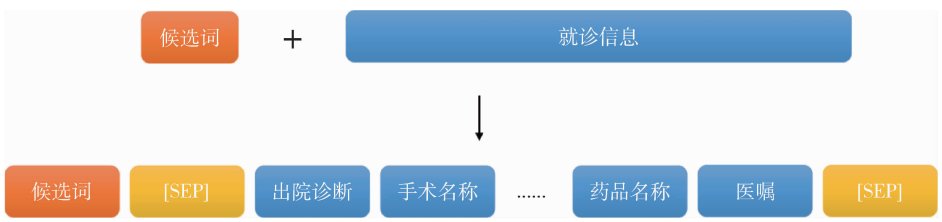

图 3 预训练模型的输入文本格式

当模型输出的标签为阳性时，表示该文本中的诊断标准词与文段相匹配，即该标准词是对应的电子病历样本的一个诊断标准词。反之，当输出标签为阴性时，则表示该标准词不是对应样本的一个诊断标准词。

5 基线模型评测结果

5.1 评测指标

本次评测任务使用平均 $F1$ 分数作为评测指标，

相关指标计算方法如下：

$$精确率 = \frac{预测出的真实标准词数量}{预测标准词数量} \quad (1)$$

$$召回率 = \frac{预测出的真实标准词数量}{真实标准词数量} \quad (2)$$

$$F1\ 分数 = \frac{2}{\frac{1}{精确率} + \frac{1}{召回率}} = \frac{2 \times 精确率 \times 召回率}{精确率 + 召回率} \quad (3)$$

5.2 实验设置

实验环境包括 Tesla V100 GPU、16G 显存、PyTorch 1.11.0 深度学习框架。在候选模块中，对每个样本选出 25 个重要程度最高的标准词作为候选集。在匹配模块中，对预训练模型微调时采用批数据训练，将每批数据量设置为 2，迭代次数设置为 10，学习率设置为 1e－6。

5.3 实验结果与分析

5.3.1 *实验结果* 由于数据集中就诊信息类别较多，实验控制了输入模型的就诊信息文本构成，并对不同输入条件下的实验结果进行对比。结果显示当输入模型文本构成是除“药品”和“医嘱”以外的其他所有就诊信息时，模型性能最优。此时，候选模块中真实标准词被选中的比例最高，为 64.382%。在训练过程中，如果不对候选模块产生的候选集进行其他操作，模型最终的 *F*1 分数为0.395 31。如果将没有被候选模块选入候选集的真实标准词也额外添加进候选集，则模型最终的 *F*1 分数为 0.460 86，提高了 16.582%，见表 3。

表 3 实验结果

是否在训练集中添加真实标准词到候选集	输入模型的文本构成	真实标准词被选入候选集的比例（%）	精确率	召回率	*F*1 分数
否	所有就诊信息	44.316	0.273 40	0.237 07	0.253 94
是	所有就诊信息	44.316	0.342 18	0.310 21	0.325 41
否	入院诊断＋出院诊断	61.968	0.361 04	0.401 19	0.380 06
是	入院诊断＋出院诊断	61.969	0.427 36	0.459 74	0.442 96
否	仅医嘱	31.281	0.063 22	0.040 92	0.049 68
是	仅医嘱	31.281	0.106 72	0.098 94	0.102 68
否	除医嘱外的其他所有就诊信息	63.912	0.372 11	0.400 86	0.385 95
是	除医嘱外的其他所有就诊信息	63.912	0.448 41	0.467 52	0.457 77
否	除药品和医嘱外的其他所有就诊信息	64.382	0.388 19	0.402 70	0.395 31
是	除药品和医嘱外的其他所有就诊信息	64.382	0.452 23	0.469 83	0.460 86

5.3.2 *结果分析* 根据前文表 2 的统计数据，数据集的“医嘱”和“药品”的文本长度远超其他就诊信息，而其本身蕴含的与真实标准词相关的信息却非常少，尤其是“医嘱”信息。例如在一条数据集样例样本中，医嘱为“11pm 后禁水禁食［嘱托］，一次性使用吸氧管 ot－mi－100 型小号（零感），一次性使用无菌注射器带针 20ml（kdl/康德莱）…”，药品名称为“中/长链脂肪乳（c8－24），亚甲蓝，氯化钠，艾司唑仑，艾普拉唑…”，这两类信息基本不包含与医学诊断相关的概念，也就无法分配医学诊断标准词。同时，当输入文本去除“医嘱”和“药品”信息后，由于输入的文本长度大幅降低，TF－IDF 算法中的词频权重（TF 值）将会相对增高，使得算法对不同标准词有更高的区分度，从而使候选模块能达到更好的选择效果，并影响后续匹配模块的性能。此外，词表中有对同一病症不同程度的症状描述，如高血压、高血压病 1 级（低危）、高血压病 2 级（中危）等，导致 TF－IDF 算法在这种情况下不能很好地选择出正确的候选标准词，匹配模块无法将样本与真实的诊断标准词进行匹配，从而无法发挥出最好的模型性能。另一方面，由于训练样本数量较少，匹配模块中经过微调

后的预训练模型也难以达到预期的分类效果。这反映出基线模型还存在一定改进空间，可以改进的方向：一是目前基线模型的候选模块仅使用 TF－IDF 算法来挖掘就诊信息中的词语与标准词之间统计意义上的相似度，可以考虑计算它们之间的语义相似度，并将二者加权作为新的排序权重；二是在匹配模块可以尝试使用其他预训练模型，例如用于生物医学文本挖掘的双向编码器（Bidirectional Encoder Representations from Transformers for Biomedical Text Mining，BioBERT）[9]，该模型在面向生物医学信息文本时能得到更好的预测效果。

6 结语

本文介绍了临床诊断编码评测任务，详细阐述评测所使用数据集来源和组成以及基线模型的构建思路。基线模型由候选和匹配两个模块构成，候选模块通过 TF－IDF 算法将每个样本可能的诊断标准词选择出来构成候选集，匹配模块通过预训练模型将每个候选标准词和病历样本中就诊信息进行匹配，若成功匹配则该候选标准词成为样本中一个预测的诊断标准词。基线模型在测试集上的最终 *F*1 分数为 0.460 86，性能有待改进。自然语言处理技术在医学领域有相当重要的应用价值，在临床诊断编码场景中，利用电子病历对应的诊断标准词，医疗工作者能更好地管理病历数据，提高医疗工作效率。临床诊断编码数据集将为未来相关研究的开展提供重要支持。

参考文献

1 Sun W, Cai Z, Li Y, et al. Data Processing and Text Mining Technologies on Electronic Medical Records: a Review [EB/OL]. [2022－05－26]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5911323/pdf/JHE2018－4302425.pdf.

2 胡海洋，赵从朴，马琏，等. 基于膨胀卷积神经网络的中文医疗命名实体识别研究 [J]. 医学信息学杂志，2021，42（9）：39－44.

3 孙曰君，刘智强，杨志豪，等. 基于 BERT 的临床术语标准化 [J]. 中文信息学报，2021，35（4）：75－82.

4 翁志雄，余志金，林燕峰，等. 基于电子病历的临床诊断编码分值库的建立和应用研究 [J]. 中国当代医药，2020，27（19）：196－199.

5 World Health Organization. The ICD－10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines [M]. Geneva: World Health Organization, 1992.

6 Sung M, Jeon H, Lee J, et al. Biomedical Entity Representations with Synonym Marginalization [EB/OL]. [2022－05－26]. https://arxiv.org/pdf/2005.00239.pdf.

7 Ramos J. Using TF－IDF to Determine Word Relevance in Document Queries [C]. Washington DC: The First Instructional Conference on Machine Learning, 2003.

8 Devlin J, Chang M W, Lee K, et al. BERT: Pre－training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. [2022－06－30]. https://arxiv.org/pdf/1810.04805v1.pdf.

9 Lee J, Yoon W, Kim S, et al. BioBERT: a Pre－trained Biomedical Language Representation Model for Biomedical Text Mining [J]. Bioinformatics, 2020, 36（4）：1234－1240.