# 基于预训练语言模型的公众健康问句分类

谢甲琦 李 政

（中国人民解放军总医院第八医学中心心内科 北京 100091） （大连莘火科技有限公司 大连 116023）

〔摘要〕 以“公众健康问句分类”任务算法评测大赛参赛项目为例，阐述通过对开源标注数据进行训练提升模型对公众健康问句识别能力的方法，包括模型构建方法、数据描述与预处理、模型算法等，为相关研究提供参考。

〔关键词〕 公众健康；问句分类；深度学习；预训练语言模型

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2021.12.006

**Classification of Public Health Questions Based on Pre－trained Language Model** *XIE Jiaqi, Department of Cardiology, the 8th Medical Center of Chinese PLA General Hospital, Beijing 100091, China; LI Zheng, Dalian Shenhuo Technology Co. Ltd., Dalian 116023, China*

〔**Abstract**〕 Taking the competition event of “Classification of Public Health Questions” algorithm evaluation contest as an example, the paper presents a method to improve the model's ability to identify public health questions by training open－source annotated data, expounds the model construction method, data description and preprocessing, model algorithm and so on, and provides references for related study.

〔**Keywords**〕 public health; question classification; deep learning; pre－trained language model

## 1 引言

近年来随着线上诊疗的普及，基于大数据、数据挖掘、人工智能等技术对医疗数据进行智能化处理并应用于辅助医生诊疗决策，提高了疾病早期识别率以及医疗质量与效率[1－2]。医学数据挖掘算法是发挥医学数据价值的重要技术手段，然而现存互联网医疗数据通常为自然语言记录，具有碎片化、非标准化特征，无法提供足够标注数据进行学习和训练。本研究拟通过对开源标注数据进行训练，提升模型对公众健康问句的识别能力。

〔修回日期〕 2021－05－21

〔作者简介〕 谢甲琦，博士，主治医师，发表论文 5 篇，参编专著 1 部；通讯作者：李政。

## 2 研究背景

### 2.1 概述

智慧医疗中一个重要场景就是智能识别公众在互联网提问的问句类别。在中华医学会医学信息学分会举办的“公众健康问句分类”任务算法评测大赛中，主办方提供已标注公众健康问句的数据，参赛者需要构造公众健康问句分类模型。训练深度学习模型需要大量人工标注数据，而此次赛事提供的已标注数据量较少，无法训练效果较好的深度学习模型。

### 2.2 模型构建

受到近两年预训练语言模型的启发，尝试采用预训练语言模型[3－5]（如 BERT 等）基于已有训练

数据进行微调，得到较理想结果。同时发现由于微调数据量较少，预测效果较不稳定，不利于模型未来在实际医疗场景中的应用。受到对抗训练研究启发，在微调模型的同时采用对抗训练策略，使得模型稳定性得到增强。从观察数据中发现部分类别样本数量较少，较难学习到判断少类模型的模式。由此将深度学习模型与人工先验知识结合，提高少量类别的识别能力。

### 2.3 创新点

采用多种预训练语言模型，基于已标注的公众健康问句数据进行微调，显著提高公众健康问句分类模型的学习能力；使用对抗训练来提高模型预测稳定性；充分结合深度学习模型和人工先验知识，提高模型对数量较少样本的判别能力。

## 3 数据描述与预处理

### 3.1 数据来源

数据来源于由 Qcorp 官网收集标注的中文公共健康问句数据库[6]，共有 5 000 条标注后的语句及 7 101 个标签，其标注的类别为第 1 层上 7 个广义的类别及第 2 层上的 29 个主题类别，这些标注全部由 8 位标注员共同完成。比赛数据使用以上所标注的 5 000 条语句以及标注的第 1 层的 6 个类别。6 个类别分别是 category_ A 诊断、category_ B 治疗、category_ C 解剖及生理学、category_ D 流行病学、category_ E 健康生活方式、category_ F 择医。

### 3.2 数据描述

问句描述大多与内科、外科、妇产科、儿科、传染病和中医等科目有关，但分布不均衡，更多分布于内科、感染、儿科、妇产科生殖领域，外科和中医科因为通常需要医生诊室检查而线上数据较少，此种分布贴合实际。近 1/3 的问句标注为多个，显示出患者倾向于一次询问多个问题，见表 1。例如 QC0000000001 描述的问题与“三高”治疗有关，所以第 1 层标注治疗（B），治疗询问阿司匹林、血塞通分散片的选择问题及其他推荐药物，第 2 层标注为药物选择（B02）。QC1000002982 是询问右心室肥大的临床意义，归类为检查结果解读（A01）、是否可以治愈、询问预后（D03）。

**表 1 样本实例**

| ID | 问句 |
|---|---|
| QC0000000001 | 病情描述：患者具有典型“三高”症状，想吃拜阿司匹林做为预防用药，但是出现过敏症状。曾经治疗情况和效果：血压高，3 年前检查出糖尿病。长期服用二甲双胍和降压药。想得到怎样的帮助：患者具有典型“三高”症状，想吃拜阿司匹林做为预防用药，但是出现过敏症状。请问可以用血塞通分散片替代白阿司匹林做为预防性用药，长期服用吗？另外还有其他药推荐吗？ |
| QC1000002982 | “去医院检查，结果是右心室肥大，有什么危险？可以治愈吗？” |

### 3.3 数据预处理

样本问句中存在 3 种 html 标签，分别是 < br/ > 、< div > 、< img > 。< br/ > 是常见的换行符标签，在训练集中还有 < BR > 等形式。共 42 条问句存在此种换行符标签。存在 < div > 标签的问句数量为 9 个，< img > 标签为 3 个。这种 html 标签并没有文字上的语义，代表特殊含义存在，如 < div > 标签后基本都有医生回答，说明句子中出现 < div > 标签都是带有回答的问句。因此在本次比赛中保留原有问句语义信息，未对样本中 html 标签进行剔除。

### 3.4 数据分布

训练集中 6 个类别数量不平衡。其中 category_ 的正样本数为 1，category_ F 正样本数为 227，category_ A 正样本最多为 1 858。问题长度大多不超过 500，最小长度为 4，最大长度为 2 093。测试集样本数量为 3 000 条，最终任务需要对测试集问句进行分类判断。每个样本可能存在 1 个或多个以上标签分类。评判标准采用 Macro F1 score 作为评判健康

问句分类效果标准，Macro F1 score 具体计算方式可以参考 Scikit - Learn 的官方文档[7]。最终线上测试集的结果为 6 个标签分类 Macro F1 score 平均分数。

## 4 模型算法

### 4.1 预训练语言模型微调

预训练语言模型是采用无监督方法对超大语言模型进行训练的模型。其代表模型有 BERT、BERT - wwm 和 RoBERTa 等。基准方案基于预训练模型进行公共健康问句分类任务，见图 1。

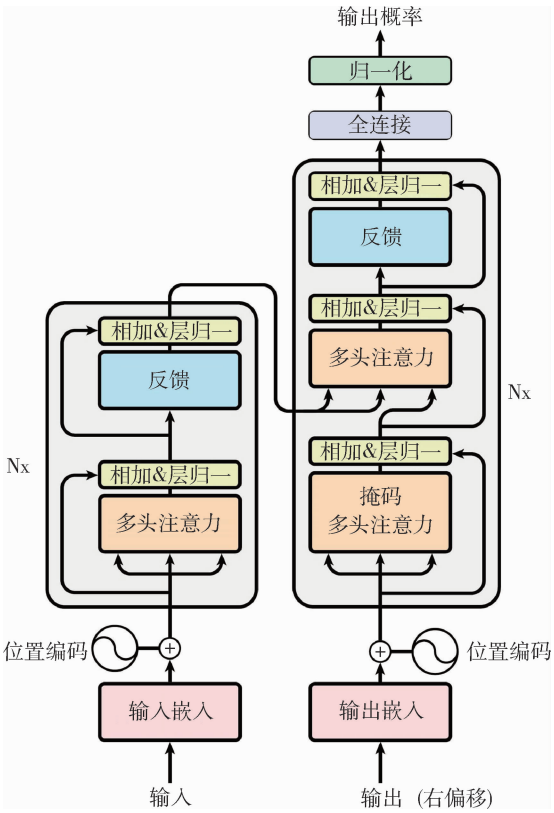


图 1 预训练模型中采用的 Transformer 网络

左侧为编码器，包含多头注意力模块和 1 个全连接模块，用于帮助输入语句转为学习表征向量；右侧为编码器，连接左侧编码器输出及产生预测的概率值。用于 Transformers 结构的输入向量来自词向量、句子标记向量以及掩码标记向量，3 部分向量相加进入到 Transformers 输入端，见图 2。针对中文预训练模型，选择哈工大讯飞联合实验室发布的 BERT - wwm 和 RoBERTa - wwm - ext - large。一方面，考虑到传统自然语言处理（Natural Language Processing，NLP）中的中文分词，应用全词覆盖（Whole Word Masking，WWM）方法，使用中文维基百科（包括简体和繁体）进行训练，对组成同一个词的汉字全部进行覆盖。另一方面，RoBERTa - wwm - ext 结合中文 WWM 技术以及 RoBERTa 模型优势以获得更好实验效果。

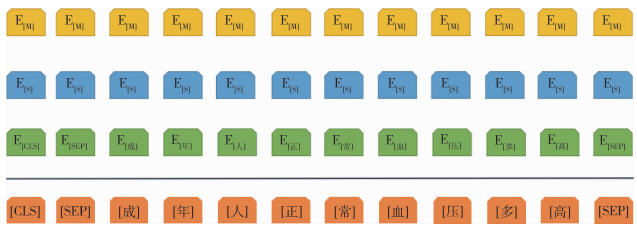


图 2 预训练模型的输入格式

### 4.2 基于先验知识的规则挖掘

category_ C 在训练集中的个数为 1，模型在训练过程中难以学习有效知识。可以通过制定规则的方式对 category_ C 的挖掘进行有效预测。例如询问组织、器官、系统、人体代谢或解剖结构、生理过程的问题定义为 C，“QC1000002327，有哪位医生可以告诉我髋关节在哪个位置？最近减肥做深蹲后膝盖酸疼”，规则为包含脏器名称“髋关节”及询问词语“哪个位置?”，标注为 category C。基于此模型纳入常见解剖及生理名称，如排卵期、血液循环等，在成为问句时即定义为 category C。这种方法虽然有助于提升实验结果，但精确度有限，在实际项目应用中可作为辅助判断依据。

## 5 实验结果

### 5.1 实验设置

5.1.1 实验设备及参数 实验环境为带有 V100 及 GTX 1080 显卡、深度学习框架为 Torch1.5.0 以上环境的服务器。实验中损失函数采用 Torch 自带的 BCEWithLogitsLoss 函数，用于学习多标签任务。

5.1.2 K - Fold 交叉验证 K - Fold 是一种交叉验证方法。将训练集分成 $K$ 份，其中 1 份为验证集，其他为新训练集。每个模型都可以预测测试

集，最后求平均加权，有助于提升模型拟合效果，见图 3。

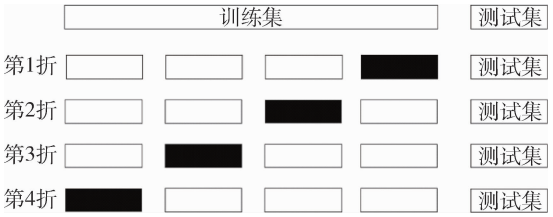


图 3 四折交叉验证数据划分

### 5.2 预训练语言模型的微调方法

实验中以 Bert - wwm 和 Robetra - large 预训练模型为选择模型做五折实验，通过改变提取问句的长度，分别对验证数据的每一类（除 category_ C）做 F1 - score 分数平均统计，每次实验需要加载预训练模型，学习速率采用默认值 1e - 5，每隔 100 次迭代进行 1 次验证结果，迭代次数均在 4 000 次内完成，保存验证 F1 - score 最好的分数模型权重为最终每折模型权重。

### 5.3 微调实验结果

Bert_ wwm_ 400（Bert - wwm 模型及截取文本长度为 400）在 category_ A、category_ B、category_ D、category_ E、category_ F 的结果分别是 0.848 7、0.927 0、0.704 1、0.783 9、0.793 0。Bert_ wwm _ 512 的实验结果为 0.848 4、0.929 9、0.696 1、0.799 7、0.808 9。可知截取文本长度超过 400 以后对于 category_ A、category_ B 较 category_ D、category_ E、category_ F 影响不大。Roberta - lagre 实验数据均比 Bert - wwm 提高 2% 以上，Roberta - lagre_ 512 实验数据为 0.873 8、0.938 1、0.745 6、0.822 2、0.847 9。

### 5.4 规则实验

根据预训练模型的微调实验结果，微调结果对验证集 category_ C 分类没有帮助，因此线上测试集较难找出 category_ C 分类样本。根据专家的建议，采用一定规则找出可能是 category_ C 标签的困难样本，如提取问句中存在关键字为（危险期 or 安全期 or 排卵期）以及（天 or 日 or 时间 or 周 or 星期）来判断是 category_ C 分类的依据。通过规则提取找出样本数为 26 条，判断正确数目为 3 条，精确度不高，最终对线上结果提升 2%。

### 5.5 线上线下分数对比

最终实验结果以线上分数为准，经线上线下分数对比可知模型截取问句长度当超过 400 之后对性能影响不大，为节省较多计算资源最终选择 400 长度；Roberta - large 参数远大于 Bert - wwm，其效果较好说明选择适当预训练语言模型具有积极意义；在加入人工先验知识规则后，Roberta - large 效果进一步提高，说明人工规则对挖掘少样本具有较好的补充作用，见表 2。

表 2 线上线下分数对比

| 模型 | 长度 | 有无对抗 | 验证集 F1 score | 测试集 F1 score |
|---|---|---|---|---|
| Bert - wwm | 400 | 无 | 0.676 1 | 0.666 2 |
| Bert - wwm | 512 | 无 | 0.680 5 | 0.666 7 |
| Roberta - large | 400 | 无 | 0.699 7 | 0.671 5 |
| Roberta - large | 512 | 无 | 0.704 6 | 0.670 1 |
| Roberta - large | 400 | 有 | 0.698 7 | 0.669 9 |
| Roberta - large + rule | 400 | 无 | 0.699 7 | 0.691 8 |

## 6 结语

本文主要解决公众健康问句分类问题，在训练数据不充分情况下合理选择以 Bert - wwm 和 Robetra - large 为主的预训练模型为选择模型做五折实验并进行微调，取得较好效果，同时加入人工规则方案，进一步提高模型性能，最终方案 Macro F1 分数为 0.691 9（6/396），能够为项目提供一定辅助判断，但方案在对困难标签 category_ C 的分类方面尚存在不足，有待进一步改进。

## 参考文献

1 Wang F Y. Parallel Healthcare: Robotic Medical and Health Process Automation for Secured and Smart Social Healthcares [J]. IEEE Transactions on Computational Social Systems, 2020, 7 (3): 581 - 586.

（下转第 43 页）

本体进行了技术层面的质量评估，但考虑到临床实践环节对数据真实性、信息准确性、响应及时性需求，该本体对临床的辅助程度仍需要对其内容全面性和真实世界可应用性进行验证；同时还需要拓宽黑色素瘤本体的应用场景，将本体嵌入临床决策支持系统及知识发现与推荐系统中，为医生、患者提供更具象化、更深入的知识服务。

## 参考文献


1 赵洁，司莉．国内外生物医学领域本体研究与实践进展［J］．数字图书馆论坛，2020（8）：7－14.

2 于凡，雷行云，高星，等．基于临床指南的糖尿病本体构建及语义检索模型设计［J］．医学信息学杂志，2018，39（5）：45－50.

3 张莉，王玉廷．基于 Protege 的糖尿病本体构建［J］．科学咨询（教育科研），2020（3）：9－11.

4 刘智锋，夏晨曦，黄梨，等．糖尿病领域本体构建与语义推理实现［J］．中华医学图书情报杂志，2017，26（9）：7－11.

5 刘何心．糖尿病本体的构建与检索研究［D］．长春：吉林大学，2014.

6 洪亮，石晓月．医学本体构建方法研究——以脑区与自闭症为例［J］．信息资源管理学报，2020，10（2）：80－90.

7 El－Sappagh S，Ali F．DDO：a diabetes mellitus diagnosis ontology［J］．Applied Informatics，2016，3（1）：5.

8 El－Sappagh S，Kwak D，Ali F，et al．DMTO：a realistic ontology for standard diabetes mellitus treatment［J］．Journal of Biomedical Semantics，2018，9（1）：8.

9 Özgü C，Emine S，Murat O，et al．Implementing Thyroid Ontology for Diagnosis and Care［C］．Barcelona：The Sixth International Conference on Global Health Challenges，2017.

10 岳丽欣，刘文云．国内外领域本体构建方法的比较研究［J］．情报理论与实践，2016，39（8）：119－125.

11 尚新丽．国外本体构建方法比较分析［J］．图书情报工作，2012，56（4）：116－119.

12 Noy N F．Ontology Development 101：A Guide to Creating Your First Ontology：Knowledge Systems Laboratory，Stanford University［EB/OL］．［2021－05－28］．http：//www.ksl.stanford.edu/people/dlm/papers/ontology－tutorial－noy－mcguinness.pdf.

13 王向前，桂冬冬，李慧宗．面向文本的本体自动构建研究综述［J］．图书馆理论与实践，2019（4）：45－50.

14 Mazen A，Mahmood M K，Susan S．Linked Open Data－based Framework for Automatic Biomedical Ontology Generation［J］．BMC Bioinformatics，2018，19（1）：319－332.

15 《临床肿瘤学杂志》编辑部．中国黑色素瘤诊治指南（2011版）［J］．临床肿瘤学杂志，2012，17（2）：159－171.

16 赵辨．中国临床皮肤病学［M］．南京：江苏科技技术出版社，2017.

17 Mayo Clinic．Melanoma－Symptoms and Causes［EB/OL］．［2021－03－10］．https：//www.mayoclinic.org/diseases－conditions/melanoma/symptoms－causes/syc－20374884.

18 百度百科．黑色素瘤［EB/OL］．［2021－03－10］．https：//baike.baidu.com/item/%E9%BB%91%E8%89%B2%E7%B4%A0%E7%98%A4/507672？fr=aladdin.

19 袁凯琦，邓扬，陈道源，等．医学知识图谱构建技术与研究进展［J］．计算机应用研究，2018，35（7）：1929－1936.

20 The National Center for Biomedical Ontology．Prostate Cancer Ontology［EB/OL］．［2021－05－28］．https：//bioportal.bioontology.org/ontologies/PCAO/？p=summary.

21 The National Center for Biomedical Ontology．Thyroid Cancer Ontology［EB/OL］．［2021－05－28］．https：//bioportal.bioontology.org/ontologies/TCO/？p=summary.


（上接第 36 页）


2 Fraust A，Olive F．Preface of Virtual Special Issue on Smart Pattern Recognition for Medical Informatics［J］．Pattern Recognition Letters，2019（128）：146－147.

3 Jacob D，Chang M W，Lee K，et al．BERT：Pre－training of Deep Bidirectional Transformers for Language Understanding［J］．NAACL－HLT，2019（1）：4171－4186.

4 Cui Y，Che W，Liu T，et al．Pre－Training with Whole Word Masking for Chinese BERT［EB/OL］．［2019－06－30］．https：//www.researchgate.net/publication/333892280_Pre－Training_with_Whole_Word_Masking_for_Chinese_BERT.

5 Liu Y，Ott M，Goyal N，et al．A Robustly Optimized BERT Pretraining Approach［EB/OL］．［2019－07－30］．https：//www.researchgate.net/publication/334735779_RoBERTa_A_Robustly_Optimized_BERT_Pretraining_Approach.

6 Guo H H，Na X，Li J．Qcorp：An Annotated Classification Corpus of Chinese Health Questions［J］．BMC Medical Informatics and Decision Making，2018，18（1）：16.

7 Pedregosa F，Varoquaux G，Gramfort A，et al．Scikit－learn：Machine Learning in Python［J］．JMLR，2011（12）：2825－2830.