基于信息增强 BERT 的阴阳性判别

杨 旋

（浙江大学医学院附属妇产科医院 杭州 310003）

薛 敏

（华东理工大学 上海 200237）

王 翔

（杭州祺鲸科技有限公司 杭州 311215）

沈晨杰

（杭州电子科技大学 杭州 310018）

〔摘要〕 介绍临床发现阴阳性判别任务要求，提出一种基于临床发现及其上下文信息增强 BERT 的阴阳性判别方法，阐述总体思路和建设路径，分析实验结果，该方法在 2021 年度中国健康信息处理大会（CHIP 2021）医学对话临床发现阴阳性判别任务的测试集上模型集成 Macro – *F*1 值达 78.1%。

〔关键词〕 在线问诊；阴阳性判别；预训练语言模型；人工智能；自然语言处理

〔中图分类号〕R – 058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673 – 6036.2023.05.014

Classifying Positive and Negative Based on Information Enhanced BERT *YANG Xuan, Women's Hospital School of Medicine, Zhejiang University, Hangzhou 310003, China; XUE Min, East China University of Science and Technology, Shanghai 200237, China; WANG Xiang, Hangzhou Qijing Technology Co. Ltd., Hangzhou 311215, China; SHEN Chenjie, Hangzhou Dianzi University, Hangzhou 310018, China*

〔**Abstract**〕 The paper introduces the task requirements of classifying positive and negative clinical findings, proposes a positive and negative discrimination method based on clinical findings and its context information enhanced BERT, expounds the general idea and construction path, and analyzes the experimental results. The results show that the method achieves a final Macro – *F*1 of 78.1% in the test set of the positive and negatived is crimination task of clinical findings in the 2021China Health Information Processing Conference (CHIP 2021).

〔**Keywords**〕 online consultation; classify positive and negative; pre – trained language model; artificial intelligence (AI); natural language processing (NLP)

1 引言

近年来以医患对话为主的互联网医疗需求增长迅速[1]，医生通过在线对话形式听取患者自诉、与患者交流病情，并提供相关医疗建议。然而由于优质医疗资源稀缺，医疗报告自动生成技术应用尤为重要。临床发现阴阳性判别是医疗报告生成中不可或缺的一环。2021 年度中国健康信息处理会议（China Health Information Processing Conference，CHIP）发布了一项对医学对话中临床发现进行阴阳性判别的评测任务[2]。临床发现是临床医学用于描述患者状态的概念，每个临床发现都具有明确的涵义，例如腹泻、呕吐等。阴阳性在该评测任务中是指患者主诉病情描述和医生诊断

〔修回日期〕 2023 – 04 – 24

〔作者简介〕 杨旋，硕士；通信作者：王翔，硕士。

判别中的阴性和阳性，即判断某一种明确的临床发现是否为患者患有，或医生推断患者当前或将来患有。

本文参考情感分析思路，将阴阳性判别任务转化为分类任务。首先从医学对话中提取临床发现词的上下文语义信息，其次将临床所见词的标准化信息通过预先构建的模板加入到上下文语义信息中，最后使用微调后的预训练模型对输入进行分类。该方法在 CHIP 2021 医学对话临床发现阴阳性判别评测数据集上 Macro - *F*1 值达到 78.1%，证明该方法对阴阳性判别任务的有效性。

2 相关工作

2.1 临床发现阴阳性判别

针对医学对话临床发现阴阳性判别问题，以往的科研工作者仅将其视为医疗报告生成的一项必备条件，通过人工方式进行收集和整理，存在工作量大并且标注准则不一致等问题。

本文将医学对话临床发现阴阳性判别视为一项针对实体的细颗粒情感分析任务进行建模。与常规的情感分析任务不同，针对实体的细颗粒情感分析任务需要对文中提及的实体在文本所包含的情感信息分类，而不是简单地对整段文本进行情感极性分类。Tang D 等[3]借鉴问答领域中的深层记忆网络模型，结合记忆和注意力机制，引入文本中的实体信息取得当时的最优效果，说明引入实体信息对细颗粒情感分析任务的有效性，为后续针对实体的细颗粒情感分析工作打下基础。Ma D 等[4]提出利用上下文动态掩蔽或上下文动态加权建模实体的局部上下文语义信息，增强模型对于实体及其上下文语义信息的利用。Sun C 等[5]通过构建辅助句子引入实体信息，同时将针对实体的细颗粒情感分析任务转换为句子对分类任务。

2.2 BERT

双向编码器表征（bidirectional encoder representations from transformers，BERT）[6]是一种多层双向 transformer 编码器，通过多头自注意力机制建模词之间的关系，从而有效解决长期依赖问题。MacBERT[7]在 BERT 模型的基础上，使用需要被遮蔽的词的相似词作为掩码进行掩码语言模型（masked language model，MLM）任务，缓解预训练与微调阶段输入的差异。MC - BERT[8]和 Med - BERT[9]则是通过替换预训练语料，生成适应医疗领域任务的 BERT。

2.3 Prompt Learning

Prompt Learning[10]是将预训练语言模型（pre - trained language models，PLM）应用于下游自然语言任务的最新范式。在不显著改变预训练语言模型结构的情况下，使用人工或自动化构建的模板修改输入文本，从而使得 PLM 适配下游任务。

3 方法

3.1 总体思路

首先针对医学对话临床发现阴阳性判别任务的特点对数据进行预处理，并基于该数据进行任务预训练生成符合任务数据分布的预训练语言模型。然后通过在临床发现词两侧加入特殊标识符和引入临床发现词的标准化信息增强临床发现词信息，从而训练基于临床发现与上下文信息增强 BERT 的模型。最后使用投票和规则融合的方法集成模型，输出阴阳性标签。此外，通过将模型预测的阴阳性标签与数据集的原始标签做对比，筛选出更高质量的样本构成新的训练数据集，进而对模型进行迭代优化，见图 1。

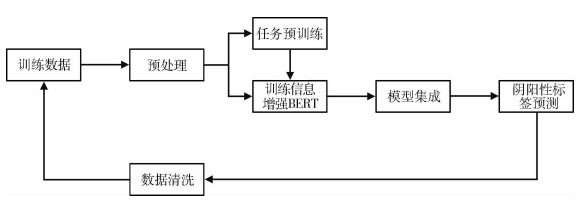

图 1 系统整体架构

3.2 数据预处理

与传统针对实体的细颗粒情感分析任务不同，

医学对话临床发现阴阳性判别任务需要考虑临床发现词在多轮对话中的上下文语义信息。因此对原始文本进行如下处理。处理 1，将文本输入者信息融入临床发现词所在的文本：若输入者为患者，则在文本前拼接“患者:”字符串；若输入者为医生，则在文本前拼接“医生:”字符串。同时，将临床发现词所在文本的下文以分号作为分隔符进行拼接，若当前临床发现词所在文本是医生输入，则拼接 3 轮下文患者输入文本；若是患者输入，则不区分下文输入者信息，直接拼接 3 轮下文输入文本。处理 2，BERT 模型输入长度存在限制，必须对输入文本操作，但常规以文本开头字符作为截断开始位置的方法容易使临床发现词无法被包含在截断后的文本中，进而导致信息缺损问题。因此使用以临床发现词为核心的截断方法，若截断末尾位置距离临床发现词结尾字符在文中的位置为 20 个字符以上，使用常规截断方法；否则以临床发现词开头字符在文中的位置减去截断长度一半的位置为截断开始位置，以临床发现词结尾字符在文中的位置加上截断长度一半的位置为截断结尾位置进行文本截断。处理 3，拼接文本长度为小于 40 个字符的上文文本。

3.3 任务预训练

开源 BERT 模型一般是基于大型通用语料进行训练，其数据分布与本次评测数据目标域不同。因此，本文使用评测训练集和测试集数据进一步对 BERT 进行预训练。仅采用 MLM 对模型进行任务预训练。MLM 通过随机使用“[MASK]”对原始输入中的部分词进行遮蔽再根据上下文信息还原建模文本的双向特征表示。具体来说 BERT 随机选择原始输入文本中 15% 的词进行替换，被选择的词有 10% 的几率被词表中随机抽样的词替换，10% 的几率保持不变，80% 的几率被替换为标签“[MASK]”。

3.4 信息增强 BERT

为了充分利用临床发现的位置信息和多轮对话上下文语义信息，对临床发现词所在文本，BERT 在文本开始和结尾位置分别插入“[CLS]”和“[SEP]”的基础上，在临床发现词前后分别插入特殊字符“[UNUSED1]”和“[UNUSED2]”，使模型可以更准确地获取临床发现词的边界信息，见图 2。

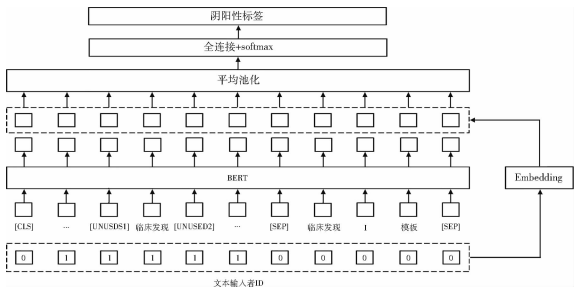

图 2 模型结构

此外，受 Prompt Learning 的启发，将临床发现词的标准化信息通过预先构建的模板加入到文本结尾，丰富模型获取的临床发现词信息。若临床发现词存在对应的标准名，则采用临床发现词 + “|标准化为” + 临床发现词所对应的标准名。若临床发现词不存在对应的标准名，则在临床发现词后拼接“|没有标准化”。

同时，本文通过构建文本输入者嵌入矩阵，生成包含文本输入者信息的特征向量，并将其与 BERT 输出文本特征向量拼接后进行平均池化，从而保证模型能够区分文本输入者信息。设置医生所输入字符对应 ID 为 1，患者所输入字符对应 ID 为 2，其他字符对应 ID 为 0，图 2 中展示了文本皆是由医生所输入时的情况。

3.5 模型集成

在提交结果时使用集成模型，按上述模型设定，分别使用 MC - BERT、Med - BERT、MacBERT - Large 以及基于任务预训练后 Mac - BERT - Large 共 4 种预训练模型进行微调训练。每种预训练模型均采用 10 折交叉验证的方法训练 10 组模型。使用投票法对每种预训练模型通过交叉验证得到的 10 组模型进行集成，获取最终的输出结果。由于本次评测任务数据集中“不标注”和“其他”的样本数据量相对较少，容易使模型将这两类样本错分为“阳性”和“阴性”。为了增加这两类标签的召回率，提出一种弱监督投票策略，即在 10 组投票结果

中，若有 2 组以上预测结果为“不标注”或“其他”，则忽略其他高票预测结果，选择该预测结果作为投票最终结果。最后，在上述集成方法融合每种预训练模型交叉验证结果的基础上，使用规则对集成融合的结果进行修正，获取最终集成融合结果。

3.6 数据清洗

对医学对话中的临床发现进行阴阳性标注需要结合上下文信息进行语义理解，因此难以保证标注的一致性。因此使用在原始训练集上训练的模型对原始训练集进行预测，过滤预测结果与原始标签不一致的样本后得到新的训练集，并使用该数据集重新进行模型训练，进而得到鲁棒性更强的阴阳性判别模型，见图 3。

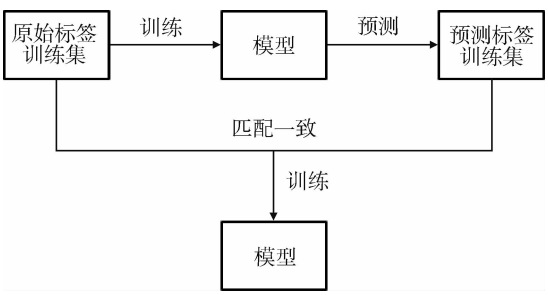

图 3 数据清洗

4 实验结果及其分析

4.1 实验结果分析

采用 BERT 的 3 种预训练模型：MC - BERT、Med - BERT 和 MacBERT - Large。其中，前两者隐藏层维度为 768，第 3 个为 1 024。3 种预训练模型采用相同的实验设置：批大小设为 64，输入文本截断长度为 200，Dropout 参数设置为 0.1，选择 AdamW 作为优化器，学习率设置为 2e - 5，训练轮数设置为 2 轮。为了减少数据不平衡对模型的影响，选用带标签平滑的交叉熵作为损失函数。此外，为了提高模型的鲁棒性，引入指数移动平均线（exponential moving average，EMA）对模型参数进行累计平均，并且采用 FGM[11] 方法向嵌入向量加入对抗扰动。本评测采用 Macro - $F1$ 作为评估指标：

$$\text{Macro} - F1 = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\frac{2 \times P_i \times R_i}{P_i \times R_i} \quad (1)$$

其中，准确率 P_i 为正确预测为类别 C_i 的样本个数/预测为 C_i 的样本个数，召回率 R_i 为正确预测为类别 C_i 的样本个数/真实 C_i 的样本个数。本次评测所提交模型的最终结果，见表 1，该结果使用的测试集为评测方提供的 B 榜测试数据，共包含 1 999 段互联网在线问诊对话。表中所列模型皆是采用投票法集成 10 折交叉验证结果后的模型。

表 1 模型 B 榜评测结果

模型	$F1$（%）
Med - BERT	76.9
MC - BERT	77.1
MacBERT - Large	77.8
MacBERT - Large + 弱监督投票	77.9
多模型集成融合	78.1

从表 1 中可以看出，采用不同的预训练模型对最终结果存在一定影响。同时，采用弱监督投票策略集成的 MacBERT - Large 预测结果相比原始投票集成的 MacBERT - Large 有一定提升，证明该策略在本次评测中的有效性。此外，评测结果也表明通过多模型集成能够较好地综合不同模型之间的优势，从而提高最终的预测效果。下面使用 A 榜测试结果进一步说明上文提及的方法和策略对于医学对话临床发现阴阳性判别的影响。评测结果，见表 2。

表 2 模型 A 榜评测结果

模型	$F1$（%）
基线	68.0
+ 数据预处理 1	72.9
+ 临床发现前后添加标识	73.9
+ 标准化信息	74.1
+ EMA	74.8
+ 数据预处理 2	75.7
+ 预训练模型替换成 MC - BERT	76.0
+ FGM	76.6
+ 预训练模型替换成 MacBERT - Large	76.9
+ 数据预处理 3	77.0
+ 交叉验证集成	77.3
+ 数据清洗	77.6

基线采用 Ernie 1.0[12] 作为预训练模型，输入文本为临床发现词所在的文本拼接上下文各一轮输入文本，批大小设为 32，输入文本截断长度为 100，其他实验设置与上文相同。

从表 2 可以看出，本文提及的方法和策略对医学对话临床发现阴阳性判别的效果提升均有一定的帮助。加入数据预处理 1 相较基础基线提升 4.9%，可以看出引入更多对话信息能显著提升模型效果。数据预处理 3 同样基于上述理念，但本文通过实验发现，大部分阴阳分类依据信息并不在上文，引入过长的上文信息只会导致更多噪声，最终选择长度为 40 的上文信息，从表 2 中可以看出，该策略可以带来一定的效果提升。在增强和丰富临床发现词信息方面，采用在临床发现词前后添加“[UNUSED1]”和“[UNUSED2]”以及增加标准化信息，前者取得了 1% 的提升，证明通过在临床发现词前后添加标识可以更多地帮助模型捕捉临床发现词的上下文语义信息。此外，数据预处理 2 通过改变截断方式，保留更多临床发现词信息，同样取得接近 1% 的提升，更进一步说明临床发现词信息对本任务的重要性。为减少训练集和测试集之间的数据分布差异，引入 EMA 和 FGM 增强模型的泛化和鲁棒性，从表 2 中可以看出这样能带来较明显的提升。同时使用 10 折交叉验证保证模型的泛化和鲁棒性。另外，本文所使用的数据清洗方法针对训练数据中错标和标签分类模糊的样本进行处理，同样可以取得一定的效果提升。

4.2 错误分析

通过对模型在验证集上的预测结果进行分析整理，归纳出导致模型预测错误的因素主要有以下 3 种。一是模型难以通过文本语义判断文本在对话中的对应关系，从而导致当对话不以一问一答形式进行时，将难以正确地识别临床发现的阴阳性。如对于“感冒”这一临床发现的阴阳性判断，患者提到的“没有”是对“流鼻涕”的否定，但模型错误地以为是对“感冒”的否定，因此将“感冒”预测成阴性，见图 3。二是由于数据集存在不平衡问题，模型倾向于将对话中出现的难以确定其阴阳性的临床发现预测为阳性，如表 3 对话里出现的乏力这一临床发现缺乏明确的上下文信息去判断其阴阳性，故应被标记为“其他”，但模型却将其预测为阳性。三是医学对话临床发现阴阳性判别任务的标注难度相对较大，导致容易出现标注不一致的问题，如表 3 中的阴道分泌物指的是一种检查，但被标注为“阳性”。

表 3 预测结果样例

对话	标注和模型预测
患者：这可能是什么问题 医生：有没有流鼻涕 医生：应该是感冒 患者：没有	“感冒”标记为阳性，模型预测为阴性
医生：有没有手心发热、口干，耳鸣，乏力健忘，易疲劳 患者：耳鸣健忘、易疲劳	“乏力”标记为其他，模型预测为阳性
医生：有的，阴道分泌物、心电图、血常规、凝血等，医生会给指导	“阴道分泌物”标记为阳性，模型预测为不标注

5 结语

医学对话临床发现阴阳性判别任务旨在判断互联网在线问诊记录中的临床发现是否为患者当前或未来大概率所患有，对自动化医疗报告生成具有重要意义。本文提出一种基于临床发现词与上下文信息增强 BERT 的医学对话临床发现阴阳性判别方法，该方法在临床发现词两侧加入特殊标识符将其定位，并使用人工构建的模板引入临床发现词的标准化信息，通过 BERT 模型获取文本特征向量后拼接文本输入者向量，融合文本输入者信息，最终输入全连接层整合信息后输出阴阳性判别结果。本文提出的方法在 CHIP 2021 医学对话临床发现阴阳性判别任务评测数据集上 Macro – $F1$ 达到 78.1%，获得该项评测第 1 名，

说明本文方法对医学对话临床发现阴阳性判别任务的有效性。未来考虑与知识图谱技术相结合，从图谱中引入问诊文本所含临床发现词的知识信息并加入模型中。

参考文献

1 冯贺霞，李韬，王佳．我国数字健康发展历程、特征及展望［J］．医学信息学杂志，2021，42（5）：9－13.

2 熊英，陈漠沙，陈清财，等．CHIP 2021 评测任务 1 概述：医学对话临床发现阴阳性判别任务［J］．医学信息学杂志，2023，44（3）：46－51.

3 TANG D，QIN B，LIU T. Aspect level sentiment classification with deep memory network［C］. Austin：Association for Computational Linguistics，2016.

4 MA D，LI S，ZHANG X，et al. Interactive attention networks for aspect－level sentiment classification［C］. Melbourne：The 26th International Joint Conference on Artificial Intelligence，2017.

5 SUN C，HUANG L，QIU X，et al. Utilizing bert for aspect－based sentiment analysis via constructing auxiliary sentence［C］. Online：Association for Computational Linguistics，2019.

6 DEVLIN J，CHANG M W，LEE K，et al. BERT：pre－training of deep bidirectional transformers for language understanding［C］. Minneapolis，Minnesota：Association for Computational Linguistics，2018.

7 CUI Y，CHE W，LIU T，et al. Revisiting pre－trained models for chinese natural language processing［C］. Online：Association for Computational Linguistics，2020.

8 ZHANG N，JIA Q，YIN K，et al. Conceptualized representation learning for chinese biomedical text mining［C］. New York：Association for Computing Machinery，2020.

9 RASMY L，XIANG Y，XIE Z，et al. Med－BERT：pretrained contextualized embeddings on large－scale structured electronic health records for disease prediction［J］. NPJ digital medicine，2021，4（1）：1－13.

10 LIU P F，YUAN W Z，FU J L，et al. Pre－train，prompt，and predict：a systematic survey of prompting methods in natural language processing［EB/OL］.［2021－07－28］. https：//arxiv. org/pdf/2107. 13586v1.

11 GOODFELLOW I，SHLENS J，SZEGEDY C. Explaining and harnessing adversarial examples［C］. San Diego：The 3rd International Conference on Learning Representations，2015.

12 SUN Y，WANG S H，LI Y K，et al. ERNIE：enhanced representation through knowledge integration［EB/OL］.［2019－04－19］. https：//arxiv. org/pdf/1904. 09.

（上接第 54 页）

3 王梓瑶，金恒江．新媒体环境下健康传播的发展研究［J］．新闻研究导刊，2023，14（1）：53－56.

4 李舒楠．融媒体视域下健康传播路径探析［J］．天中学刊，2019，34（6）：80－84.

5 匡文波，武晓立．基于微信公众号的健康传播效果评价指标体系研究［J］．国际新闻界，2019，41（1）：153－176.

6 谭盈．新冠疫情下医生自媒体健康传播效果分析——以“今日头条”为例［J］．新闻传播，2021（2）：54－55.

7 宋琼芳．自媒体时代对健康传播的启示［J］．健康教育与健康促进，2014，9（6）：474－476.

8 中国报道．抑郁症或成全球第二大疾病［J］．中国报道，2017（11）：68.

9 常军，边育红，徐欢，等．佛手治疗抑郁症的研究进展［J］．时珍国医国药，2021，32（8）：1971－1973.

10 单忠艳．根据 2018 年美国糖尿病学会标准诊断中国糖尿病患病率：全国横断面研究［J］．国际内分泌代谢杂志，2020，40（5）：314－314.

11 郭航．兼具医学知识与传播思维：后疫情时代下健康传播人才的培养路径探析［J］．视听，2020（8）：136－137.

12 魏小津．自媒体健康传播中的问题及改进策略［J］．青年记者，2019（23）：13－14.

13 张欣．抖音健康类短视频的问题及优化路径分析［J］．新媒体研究，2020，6（15）：123－125.