基于共享编码策略的医学对话阴阳性判别系统

姜逸文 伊博乐

（卫宁健康科技集团股份有限公司人工智能实验室 上海 200072）（上海交通大学密西根学院 上海 200240）

陈 旭

（卫宁健康科技集团股份有限公司人工智能实验室 上海 200072）

〔摘要〕 介绍临床发现阴阳性判别任务研究现状，提出一种基于预训练模型的临床发现阴阳性判别系统，详细阐述系统构建方法，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 77.87%。

〔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 – 6036.2023.03.009

A Shared Embedding Strategy Based System for Clinical Findings Classification in Medical Dialogues *JIANG Yiwen, AI Labs, Winning Health Technology Group Co. Ltd., Shanghai 200072, China; YI Bole, UM – SJTU Joint Institute, Shanghai Jiao Tong University, Shanghai 200240, China; CHEN Xu, AI Labs, Winning Health Technology Group Co. Ltd., Shanghai 200072, China*

〔**Abstract**〕 The paper introduces the research status of negative and positive classification tasks in clinical findings, proposes a system for clinical findings classification in medical dialogues based on the pre – trained model, elaborates the construction method of the system and analyzes the experimental results. The proposed method achieves 77.87% of Macro – *F*1 value in the 7th China health information processing conference (CHIP 2021) classifying positive and negative clinical findings in medical dialogues task.

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

1 引言

临床发现是临床医学中描述患者状态的概念集合，医学对每一个临床发现的概念都有明确的定义和说明。为了尽可能全面和精准地对患者状态进行客观描述，需要利用临床发现的概念对患者状态进行表达，其中最基础的状态就是阴性和阳性。随着在线问诊平台的兴起，患者通过在线问诊使医生了解病情、诊断疾病以获取医疗建议。由于医生资源有限，与在线问诊需求不匹配，如何通过机器辅助问诊过程、汇总患者状态描述进而实现自动化医疗

〔修回日期〕 2022 – 10 – 22

〔作者简介〕 姜逸文，算法工程师；通信作者：陈旭，博士，高级工程师，发表论文 10 篇。

问诊是一个日益重要的话题。近年来针对医患在线问诊数据，学术界展开一系列研究，例如，对医患对话文本进行手术和检查等重要信息的抽取[1]、基于对话历史自动生成医生的回复[2]等。实现自动化医疗问诊的一项基础任务是理解患者口语化描述的身体状态、明确患者临床表现的阴阳性。

临床发现阴阳性判别任务的研究在近年来主要基于深度学习[3-5]的方法。本文将临床发现阴阳性判别任务抽象为对话信息抽取问题，提出一种基于预训练语言模型的输入设计，对对话系统中的角色变化进行建模，同时提出一种基于共享编码策略的辅助序列，提示模型对提及的临床发现进行逐一预测。该方法在 2021 年中国健康信息处理大会（China Health Information Processing Conference，CHIP）任务 1（http://www.cips-chip.org.cn/2021/eval1）的评测数据集上 Macro-$F1$ 值达 77.87%。

2 研究现状

近年来针对医学对话临床发现阴阳性判别任务已有研究和探索，部分研究将临床发现识别和阴阳性判别两个任务进行统一研究，构建联合学习模型。Du N 等[3]提出基于序列标注的临床发现识别和判别联合模型。Lin X 等[4]提出一种全局注意力机制来分别获取临床发现在对话和语料集中的相关信息，通过构造症状图谱来建模不同症状之间的关系。Zhang Y 等[1]通过信息抽取的方法，利用对话交互的深层匹配模型来抽取症状及其阴阳性。与上述研究者将该任务视为分类问题不同，Li M 等[5]通过生成式的方法设计基于相对位置编码的多粒度 Transformer 模型[6]来生成症状及其标签。

针对医学对话阴阳性判别任务，本文将其抽象为对话信息抽取问题。对话信息抽取任务旨在识别对话文本中两个对象之间的关系。将患者这一对话角色视为主体，将临床发现视为客体，则医学对话阴阳性判别任务即识别主客体间关系的任务。Yu D 等[7]实验了多种基于预训练模型的输入方法，并表明一种将主客体作为辅助序列与对话文本进行拼接的方法明显优于其他输入方案。本文在此输入方法的基础上，设计一种共享编码策略，提升模型在 CHIP 2021 评测任务上的指标。采用的方法基于预训练语言模型，以双向编码器表征（bidirectional encoder representations from transformers，BERT）[8]模型结构为主。

3 任务描述

3.1 任务概述

本次评测任务旨在针对互联网医患在线问诊记录中的临床发现部分进行阴阳性的分类判别，数据来源于“春雨医生”互联网在线问诊公开数据。给定医患在线对话的完整记录以及医患交互中提及的临床发现，要求对临床发现的阴阳性类别做判断。阴阳性的定义一般认为是患者主诉病情描述和医生诊断判别中的阴性和阳性。除此之外，本次评测任务的标签还有“其他”和“不标注”。“其他”一般是指对话中一方没有回答、不知道或者回答内容不明确的情况。“不标注”通常是医生在解释一般性知识和患者当前状态条件独立不具有标注意义，以及一些在检查或者药品中出现的不予标注。

3.2 评测数据集统计分析

3.2.1 数据分布 该任务的数据中，一段完整的对话记录通常会提及多个临床发现，相同名称的提及可能由于语境的变化有不同的标签，需要联系上下文语义和对话逻辑关系做出判断。评测任务提供两份测试数据，最终以在测试集 B 上的指标为准。完整的对话文本属于长文本，由多轮对话构成，单轮对话的平均长度相对较短。评测数据分布统计，见表 1。

表 1 评测数据分布统计

数据	数据量（条）	完整对话平均字符数（个）	单轮对话平均字符数（个）	平均对话轮次（轮）	标签数量（条）
训练集	6 000	533.58	19.84	31.05	118 976
测试集 A	2 000	529.22	17.29	30.60	39 204
测试集 B	1 999	522.52	17.40	30.03	40 276

3.2.2 标签分布 分析训练集中各标签分布的统

计结果，见表 2，评测数据集存在标签分布不平衡的问题。因为在医患对话过程中，双方更倾向于关注患者已有的临床表现，医生通常会向患者解释症状和疾病相关的一般性知识，对于未出现的临床表现则不太关注，对话中一方没有回答或回答模棱两可的情况则更加少见。

表 2 标签分布统计

标签	训练集中占比（%）
阳性	62.848
阴性	11.839
其他	5.183
不标注	20.129

3.3 评测指标

该任务以 Macro – $F1$ 值作为最终评估标准，具体方法为：首先计算每个类别精确率和召回率的调和平均值（$F1$），接着对所有类别的 $F1$ 求平均值。假设任务有 n 个类别 $C_1, C_2, \cdots, C_n$，第 i 个类别的准确率和召回率分别为 P_i 和 R_i，评测任务按如下方式计算 Macro – $F1$。

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

4 方法

4.1 基于角色感知的输入序列

系统整体架构，见图 1。输入端使用基于角色感知的输入序列。不同于常规文本输入序列，对话文本中的内容通常蕴含对话者的角色信息。为了更好地对输入序列中的角色变化进行建模，基于传统 BERT 模型的输入方法，通过显式与隐式两种方式对角色的变化信息进行嵌入。显式嵌入方法通过在 BERT 模型的词表中添加额外的特殊标识符来表示每段话语的对话角色。在当前任务中，令 $[s^{(1)}]$ 表示医生，令 $[s^{(2)}]$ 表示患者，对于一段完整的对话数据 $d = u_1, u_2, \cdots, u_n$，其中 $u_i (1 \leqslant i \leqslant n)$ 表示单条对话的文本序列，n 表示总的对话轮次，那么通过将角色特殊标识符拼接在每条对话内容前可以得到 $d' = [s_1^*], u_1, [s_2^*], u_2, \cdots, [s_n^*], u_n$，作为模型的输入。隐式嵌入方法[9]通过对输入序列 d' 中的每个字添加角色嵌入向量，与 BERT 模型的字嵌入向量和位置嵌入向量等相加，得到最终的输入表示。

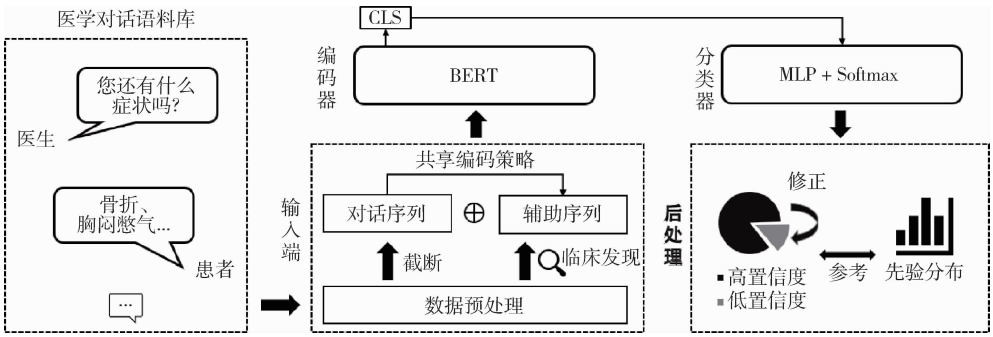

图 1 系统整体架构

4.2 基于共享编码策略的辅助序列

4.2.1 *添加辅助序列* 由于医患对话文本 d 和临床发现 $C = \{c_1, c_2, \cdots, c_k\}$ 之间是一对多关系，k 表示 d 中需要预测的样本数量。通过添加辅助序列提示模型对文本 d 中的临床发现 $c_i \in C$ 进行逐一预测。具体来说，模型的输入序列被划分为对话序列和辅助序列两部分，通过分隔符 $[SEP]$ 隔开。其中，对话序列 $d' = [s_1^*], u_1, [s_2^*], u_2, \cdots, [s_n^*], u_n$，辅助序列由临床发现 $c_i = t_1^i, t_2^i, \cdots, t_j^i$ 和患者特殊标识符 $[s^{(2)}]$ 组成，$t_\alpha^i (1 \leqslant \alpha \leqslant j)$ 表示构成临床发现的字，j 表示字符长度。辅助序列通过添加患者特殊标识符来提示模型，其预测的临床发现与患者高度相关。模型最终的输入序列为 $[CLS]$，d'，$[SEP]$，

$[s^{(2)}]$, $[SEP]$, c_i, $[SEP]$，其中［CLS］用来实现对阴阳性的分类。

4.2.2 共享编码策略 通过对数据集的观察可以发现，同一对话序列 d 中可能多次提及相同的临床发现，即：$\exists\ c_x, c_y \in C$ 且 $t_\alpha^x = t_\alpha^y (1 \leqslant \alpha \leqslant j)$。由于上下文语境不同，其标签分类结果亦不同。但是通过上述方法会产生相同的输入序列，导致歧义性，模型无法得知当前预测对象的具体位置。为解决此问题，本文提出共享编码策略，使对话序列和辅助序列中的临床发现 c_i 具有相同的输入表示。辅助序列中的临床发现 c_i，其字嵌入向量、位置嵌入向量和角色嵌入向量直接由其在对话序列中相应位置的嵌入结果共享而来。最后，通过 BERT 模型的段嵌入向量使模型能够区分两种输入序列的类型，见图 2。

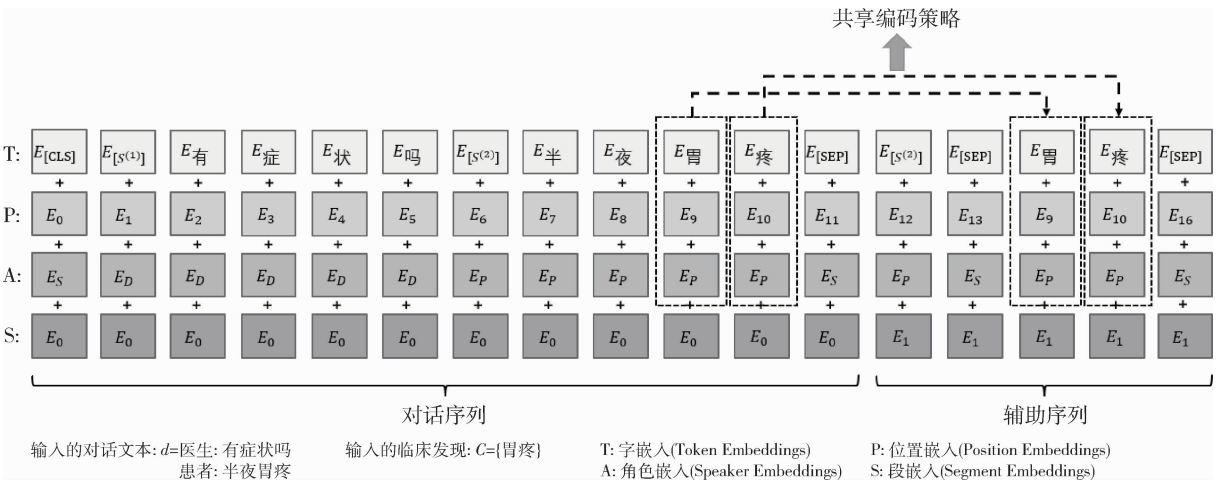

图 2 模型输入端嵌入表示

4.3 编码器和分类器

采用多种不同的中文预训练语言模型作为编码器，提取对话文本的语义特征，将编码器的结果进行解码后得到标签概率分布。本文微调的大部分中文预训练模型基于相同的 BERT 模型结构，使用不同预训练方法。假设模型的输入嵌入表示为 h_0，将 h_0 通过一个 N 层的 Transformer 网络，得到语义表示向量 h_N 作为文本的编码结果，其中 h_l 为第 l 层网络的输出。

$$h_l = Transformer(h_{l-1}), l \in [1, N] \tag{2}$$

接着模型的分类器将 $h_N[0]$，即“$[CLS]$”处的向量表示输入到多层感知机（multi-layered perceptron，MLP）中进行解码，得到 m 维向量，m 为类别数量。最后通过 Softmax 层对标签概率进行归一化，训练时使用交叉熵作为损失函数。

4.4 正则化约束方法

4.4.1 方法对比 为提高模型泛化能力，在模型训练过程中通过 R-Drop[10] 方法对模型的输出预测进行正则化约束。Dropout 是一种广泛使用的正则化技术，在模型训练过程中按照一定概率随机丢弃部分神经元，以此避免神经网络在训练过程中的过拟合。由于被丢弃的神经元是随机的，所以带有 Dropout 层的模型每次会产生不同的子模型，从而导致模型在训练和推理过程中的不一致性。为缓解此问题，本文通过 R-Drop 方法进一步对子模型的输出预测进行正则化约束。

4.4.2 R-Drop 方法 对于训练数据集 $\{(x_i, y_i)\}_{i=1}^n$，其中 x_i 是模型的输入序列，y_i 是类别标签，n 是训练集中的样本数量，模型预测的标签概率分布为 $p^\theta(y_i \mid x_i)$。对于每个训练样本 x_i，在模型中进行两次前向传播，由于 Dropout 层带来的随机性，样本经过两个不同的子模型后分别得到不同的概率分布 $p_1^\theta(y_i \mid x_i)$ 和 $p_2^\theta(y_i \mid x_i)$。R-Drop 方法采用对称的相对熵使不同子模型的输出结果尽可能一致。

$$L_i^{(KL)} = \frac{1}{2}(D_{KL}(p_1^\theta(y_i \mid x_i) \,||\, p_2^\theta(y_i \mid x_i)) +$$

$$D_{KL}(p_2^{\theta}(y_i \mid x_i) \,||\, p_1^{\theta}(y_i \mid x_i))) \tag{3}$$

结合交叉熵损失函数 $L_i^{(CE)}$，模型最终的目标函数为如下加权和，其中 α 为控制正则项的权重系数。

$$L_i = L_i^{(CE)} + \alpha L_i^{(KL)} \tag{4}$$

4.5 交替归一化后处理

通过前文表 2 的标签分布统计，发现评测数据集有标签类别不平衡的问题，这会导致模型在推理过程中存在偏差。为缓解这一问题，本文在模型推理过程中使用交替归一化[11]的后处理方法，使模型在测试集中的标签分布与训练集中的分布尽可能一致。首先，通过 TOP - K 交叉熵将模型的预测结果划分为高置信度与低置信度。假设有 m 个标签类别，模型预测的标签概率分布为 $p = [p_1, p_2, \cdots, p_m]$，对分布进行排序后选择概率最高的 k 个值，用如下公式通过 TOP - K 交叉熵计算其置信度。选定一个阈值 τ，将置信度小于 τ 的预测结果划分为高置信度，其余则为低置信度。

$$\tilde{p}_i = \frac{p_i}{\sum_{i=1}^{k} p_i} \tag{5}$$

$$H_{top-k}(p) = -\frac{1}{\log k} \cdot \sum_{i=1}^{k} \tilde{p}_i \log \tilde{p}_i \tag{6}$$

其次，利用高置信度的预测结果，逐一对低置信度的预测结果进行修正，使其分布尽可能与训练集中的先验分布一致。假设 $b_0 \in R^m$ 是模型中一个低置信度预测结果，$A_0 \in R^{n \times m}$ 是对模型 n 个高置信度预测结果进行拼接后得到的矩阵，对 A_0 与 b_0 进行拼接得到矩阵 $L_0 \in R^{(n+1) \times m}$，令 $p^{(i)}$ 为矩阵 L_0 中第 i 个样本的概率分布，那么 $p^{(n+1)} = b_0$，对 L_0 进行 γ 次交替归一化迭代后，取 $p^{(n+1)}$ 作为修正后的概率分布。交替归一化分为样本间归一化与样本内归一化。令 $p^{(r)}$ 为先验分布，直接由训练集统计得到。对于第 n 次迭代，样本间归一化如下所示，其中 $\bar{p}$ 为归一化前 L_n 的平均概率分布，该步骤旨在使 L_n 在更新后的平均概率分布等于先验分布 $p^{(r)}$。

$$L_{n'} = \frac{L_n}{\bar{p}} \cdot p^{(r)} \tag{7}$$

此时由于 $p^{(i)} \in L_{n'}$ 的概率分布之和不满足归一化要求，对每个 $p^{(i)}$ 进行样本内归一化得到更新后的 L_{n+1} 完成整个迭代。

$$p^{(i)} = \frac{p_j^{(i)}}{\sum_{j=1}^{m} p_j^{(i)}} \tag{8}$$

5 实验结果与分析

5.1 数据预处理

对于医患对话文本 d，假设给定临床发现 $C = \{c_1, c_2, \cdots, c_k\}$，$k$ 表示 d 中提及临床发现的数量，那么经过数据预处理后将生成 k 条输入序列 d' 用于训练和预测。由于 d' 为长字符序列，预训练模型的输入长度为 L，通常需要对输入的对话序列进行截断后与辅助序列进行拼接。假设 $c_i (1 \leqslant i \leqslant k)$ 的长度为 L_c，那么对话序列的最大长度为 $L - L_c - 5$（特殊标识符的数量为 5）。实验中以 c_i 为中心对 d 进行截断，使输入尽可能多地包含上下文对话信息。

5.2 实验结果

实验中通过 5 折交叉验证方法，采用多种开源中文预训练语言模型进行微调，同时采用伪标签学习训练方法，利用无标签的测试数据集来提升模型性能。根据实验结果，括号中的 A* 和 B* 分别表示由序号 1 - 2 的模型和序号 1 - 4 的模型在测试集 A 和 B 中通过集成生成的伪标签数据，见表 3。

表 3 不同预训练模型 5 折交叉验证的平均结果

序号	模型名	平均 Macro - F1（%）
1	CPT - Base	75.17
2	RoBERTa - Base	75.45
3	RoBERTa - Large（A*）	75.53
4	MacBERT - Base（A*）	75.40
5	MacBERT - Large（A*）	75.80
6	Chinese - BERT（A* + B*）	75.61

实验采用多种基于 BERT 模型结构的中文预训练的非对称语言模型。RoBERTa[12] 和 Chinese - BERT[12] 主要使用全词掩码的预训练方法。MacBERT[13] 主要通过 N - Gram 掩码和相似词替换掩码等方法进行预训练。中文预训练的非对称 Transformer 模型（Chinese pre - trained unbalanced transformer，CPT）[14] 由 1 个深层的共享编码器和 2 个分别用于理解和生成的浅层解码器构成，令理解

式和生成式解码器的特征表示为 h_E 和 h_D，将 $h_E[0]$ 即“[*CLS*]”处的表示，和 $h_D[-1]$ 即最后一个“[*SEP*]”处的表示进行拼接后输入给分类器。最终 MacBERT – Large 模型在验证集上的平均 Macro – *F*1 达到 75.80%。

5.3 模型集成

通过 5 折交叉验证，训练了 6 组模型在测试集 B 上进行集成。实验中通过两轮投票进行集成，第 1 轮投票对相同验证集上各模型的预测结果进行加权投票，权重为各模型在验证集上类别标签的 *F*1 值；第 2 轮投票对不同折上的集成结果进行硬投票，选择输出数量最多的标签作为最终集成结果。在测试集 B 中实验不同组模型的集成表现，其中下标 B 和 L 分别是模型 Base 和 Large 版本的缩写。结果表明，采用 3 组模型分别为 RoBERTa – Large、MacBERT – Large 和 Chinese – BERT 进行集成可达到最高 77.87% 的 Macro – *F*1 值，见表 4。

表 4 不同模型组合的投票结果

模型组合（单个模型的数量）	Macro – *F*1（%）
$CPT_B + Ro_{B/L} + Mac_{B/L}$（25）	77.49
$CPT_B + Ro_{B/L} + Mac_{B/L}$ + Chinese（30）	77.56
$CPT_B + Ro_{B/L}$（15）	77.63
$Ro_{(B/L)} + Mac_L$（15）	77.71
$Ro_L + Mac_L$ + Chinese（15）	77.87

5.4 实验分析

对提出的模型架构进行消融实验以测试模型中不同设置的作用。相同实验环境下采用RoBERTa – Base 模型在其中一折训练数据中进行实验，见表 5。为了更直观地观察端到端模型性能，本实验中的基准模型不包括交替归一化后处理。

表 5 消融实验结果

消融设置	Macro – *F*1（%）
基准模型	75.68
不使用基于角色感知的输入	73.18（–2.50）
不拼接辅助序列的输入	38.73（–36.95）
辅助序列不使用共享编码	70.31（–5.37）
不使用正则化约束	74.26（–1.42）
使用交替归一化后处理	75.81（+0.13）

注：括号中为与基准模型相比 Macro – *F*1 值的变化。

通过观察消融实验结果发现，辅助序列能够有效提示模型对特定临床发现进行标签预测，当辅助序列不使用共享编码策略时，单模型 Macro – *F*1 值下降 5.37%；当输入序列不包含角色变化信息时，单模型性能损失 2.50%，这表明角色感知方法和基于共享编码策略的辅助序列对本文提出的模型起到至关重要的作用。此外，正则化约束方法能提高模型的泛化能力，交替归一化后处理能对低置信度标签做出有效修正。

5.5 错误分析

5.5.1 *分析方法* 采用前文表 3 中的预训练模型在其中一折验证数据中对预测结果进行错误分析，预测结果中各类别分类状态的混淆矩阵，见表 6。观察可知：“其他”准确率偏低，主要因为样本占比较少，模型无法充分学到该类别的特征。模型在“阳性”与“不标注”之间的误分类占比最大，在阴阳性之间的误分类情况较少，证明模型能够有效判断阴阳性。

表 6 预测样本的混淆矩阵（%）

真实/预测	阳性	阴性	不标注	其他
阳性	89.86	1.15	7.67	1.32
阴性	6.61	83.39	5.04	4.96
不标注	20.14	2.59	75.18	2.09
其他	22.43	11.67	14.42	51.48

5.5.2 *分析结果* 对困难样本采样后进行错误分析，困难样本是指所有模型都预测错误的样本。结果表明有 8.24% 的样本为困难样本，其中“其他”标签中的困难样本占比最大（30.24%），“阳性”“阴性”和“不标注”标签中的困难样本占比分别为 4.96%、9.07% 和 12.16%。基于有限的医学知识进行人工分析与归纳，模型预测的错误主要分为以下 3 类。一是“其他”标签的上下文语义大多较为复杂且缺少显式提示词。不同于阴阳性标签通常可以显式地找到提示词，“其他”标签通常是对话中一方没有回答（缺少提示词），或者是需要结合上下文语义和基本常识才能推理的情况。二是模型没有对医患对话意图建模，难以判断临床发现是否与患者高度相关。三是对话内容之间存在长距离依

赖，对长文本进行截断的预处理方法会损失文本中有价值的信息，部分标签需要考虑较远的对话历史才能做出判断。

5.6 实验设置

采用相同模型参数微调不同预训练模型，最大输入序列长度为 512，分类器的 Dropout 参数设置为 0.3，批训练子集的大小为 20，选择 Adam 作为优化器，权值衰减为 0.1，学习率设置为 $1e-5$，预热比例为 10%，训练轮次设置为 5。正则化约束中权重系数 $\alpha = 4$，对于交替归一化后处理，选择概率最高的 $k(k = 2)$ 个值计算交叉熵，置信度阈值 $\tau = 0.9$，迭代次数 γ 参考单模型在验证集中的表现，一般进行 $\gamma \in \{2,3\}$ 次迭代。

6 结语

针对医学对话场景，本文提出一种基于预训练模型的临床发现阴阳性判别系统，在输入端设计基于角色感知的输入序列和基于共享编码策略的辅助序列，通过正则化约束和交替归一化后处理等方法提升系统性能。本文提出的方法在 CHIP 2021 评测任务 1 的测试集上模型集成达到 77.87% 的 Macro-$F1$ 值，单模型在验证集上达到 75.80% 的平均指标。同时本方案中对于长文本的截断预处理可能损失对话中有价值的语义信息，对临床发现进行逐一预测的方法在训练和预测时忽略了上下文各临床表现之间的关联。此外，本文使用的中文预训练语言模型并非医学对话领域的语言模型，如何利用公开的医学对话数据对语言模型进行预训练，是未来提高系统性能的重要研究方向之一。

参考文献

1 ZHANG Y, JIANG Z, ZHANG T, et al. MIE: a medical information extractor towards medical dialogues [C]. Online: The 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020.

2 LIU W, TANG J, QIN J, et al. MedDG: a large-scale medical consultation dataset for building medical dialogue system [EB/OL]. [2022-06-14]. https://arxiv.org/abs/2010.07497.

3 DU N, CHEN K, KANNAN A, et al. Extracting symptoms and their status from clinical conversations [C]. Florence: The 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019.

4 LIN X, HE X, CHEN Q, et al. Enhancing dialogue symptom diagnosis with global attention and symptom graph [C]. Hong Kong: The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (ACL), 2019.

5 LI M, XIANG L, KANG X, et al. Medical term and status generation from Chinese clinical dialogue with multi-granularity Transformer [J]. IEEE/ACM transactions on audio, speech, and language processing, 2021 (29): 3362-3374.

6 VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]. Long Beach: The 31st International Conference on Neural Information Processing Systems, 2017.

7 YU D, SUN K, CARDIE C, et al. Dialogue-based relation extraction [C]. Online: The 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020.

8 DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]. Minneapolis: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (ACL), 2019.

9 GU J C, LI T, LIU Q, et al. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots [C]. Online: The 29th ACM International Conference on Information & Knowledge Management, 2020.

10 LIANG X, WU L, LI J, et al. R-Drop: regularized dropout for neural networks [C]. Sydney: The 35th International Conference on Neural Information Processing Systems, 2021.

11 JIA M, REITER A, LIM S N, et al. When in doubt: improving classification performance with alternating normalization [C]. Punta Cana: Findings of the Association for Computational Linguistics: EMNLP (ACL), 2021.

12 CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT [J]. IEEE/ACM transactions on audio, speech, and language processing, 2021 (29): 3504-3514.

13 CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing [C]. Online: Findings of the Association for Computational Linguistics: EMNLP (ACL), 2020.

14 SHAO Y, GENG Z, LIU Y, et al. CPT: a pre-trained unbalanced transformer for both Chinese language understanding and generation [EB/OL]. [2022-06-14]. https://arxiv.org/abs/2109.05729.