CHIP 2021 评测任务 1 概述：医学对话临床发现阴阳性判别任务

熊 英 陈漠沙

（1 哈尔滨工业大学（深圳）计算机科学与技术学院 深圳 518055 （阿里巴巴 杭州 310052）
2 鹏城实验室 深圳 518055）

陈清财 汤步洲

（1 哈尔滨工业大学（深圳）计算机科学与技术学院 深圳 518055 2 鹏城实验室 深圳 518055）

〔**摘要**〕 介绍基于医学对话的信息抽取相关研究现状，结合第七届中国健康信息处理会议组织的基于医学对话的临床发现阴阳性判别评测任务，阐述该任务评测数据集、评估指标，对比介绍前 4 名参赛者的方法并进行总结，以推进医疗人工智能领域的发展。

〔**关键词**〕 中国健康信息处理会议；在线问诊；阴阳性判别；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.1673－6036.2023.03.008

Overview of the CHIP 2021 Shared Task 1: Classifying Positive and Negative Clinical Findings in Medical Dialog *XIONG Ying*, *1School of Computer Science and Technology*, *Harbin Institute of Technology (Shenzhen)*, *Shenzhen 518055*, *2Peng Cheng Laboratory*, *Shenzhen 518055*, *China*; *CHEN Mosha*, *Alibaba Group*, *Hangzhou 310052*, *China*; *CHEN Qingcai*, *TANG Buzhou*, *1School of Computer Science and Technology*, *Harbin Institute of Technology (Shenzhen)*, *Shenzhen 518055*, *2Peng Cheng Laboratory*, *Shenzhen 518055*, *China*

〔**Abstract**〕 The paper introduces the research status of information extraction based on medical dialogue, combined with a medical conversation－based clinical finding negative and positive discrimination task organized by the 7th China Health Information Processing Conference (CHIP 2021), expounds the evaluation data set and evaluation indicators of the task, compares and summarizes the effective methods of the top 4 contestants, so as to promote the development of medical artificial intelligence (AI) field.

〔**Keywords**〕 China Health Information Processing Conference (CHIP); online consultation; positive and negative classification; artificial intelligence (AI); natural language processing (NLP)

1 引言

随着互联网的普及，“互联网＋医疗”模式发展迅速，国家在“十四五”规划建议中明确提出要支持社会办医，推广远程医疗。这一系列因素促使

〔修回日期〕 2022－10－22
〔作者简介〕 熊英，博士，发表论文 6 篇。

在线问诊平台迅速兴起与发展，如“丁香医生”“春雨医生”“好大夫在线”“平安好医生”等。越来越多的患者通过在线问诊平台与医生进行交流，得到专业医生的诊断与建议，节约出行和排队等时间成本。在问诊过程中，患者会对自身情况进行口语化的描述，医生通过听取患者描述进行针对性问诊并对患者的主诉进行补充。通过这种交互方式，医生能够初步了解患者的基本情况，进而对患者进行诊断、提供相关医疗建议。

医学在线问诊平台逐渐成为刚需，然而经验丰富的医生数量有限，且医生很难挤出大量时间进行在线问诊。因此，通过采用自然语言处理（natural language processing，NLP）技术自动挖掘医学对话中的信息有助于节约专业医生诊断时间。在临床医学中，临床发现需要抽取的关键信息，主要是指患者状态描述的概念集合。每一个临床发现概念具有明确的涵义、定义与说明，如腹泻、呕吐等。为了尽可能全面和精准地对患者状态进行客观描述，需要利用严谨的临床发现概念对患者状态进行表达，其中最基本的状态就是阴性和阳性，也就是患者是否存在或者发生某一种明确的临床发现。判别临床发现阴阳性需要考虑对话上下文的联系以及逻辑关系，对临床发现是否是“阳性”“阴性”“其他”或“不标注”进行分类。“其他”一般是指用户没有回答，或是回答不明确、不知道。“不标注”则是无意义的，与患者目前状态相独立的，例如医生解释的一般知识。数据集标注样例，见表 1。

为了促进医疗信息处理社区更广泛地研究临床医学对话中临床发现的阴阳性问题，2021 年度中国健康信息处理会议（China health information processing conference，CHIP）创建了医学对话临床发现阴阳性判断任务。该任务提供了一个公开评测平台供大家评估模型与算法，最终以 Macro – $F1$ 作为评测指标。评测过程分为 A 榜和 B 榜排名，A 榜共 81 支队伍提交结果，其中最高结果 Macro – $F1$ 值为 78.03；B 榜共 15 支队伍提交结果，其中最高结果 Macro – $F1$ 值为 78.10。文本将对 CHIP 2021 任务 1 中的评测数据以及前 4 名评测结果进行分析和总结。

表 1 CHIP 2021 评测任务 1 数据集标注

样例	对话	临床发现与标注
1	患者：医生您好，从昨天晚上开始肚子一直疼，吃了布洛芬有所缓解。	肚子一直疼——阳性
	医生：肚子疼，是上腹部疼么？	肚子疼——阳性
		上腹部疼——阴性
	患者：不是，主要是下腹部疼。	下腹部疼——阳性
	医生：是针扎样的疼么？	针扎样的疼——其他
	患者：不知道，描述不出来，有点抽筋的那种疼。	抽筋的那种疼——其他
	医生：这种情况考虑为急性肠胃炎导致的，急性肠胃炎可能除了腹疼之外，可能还会引起腹泻等，需要及时补充水分。	腹疼——不标注 腹泻——不标注
2	患者：坐起来就不怎么痛①，躺着就痛②，站着不动也不怎么痛③，走路慢点也还好，快点就痛④	痛①——阴性 痛②——阳性 痛③——阴性 痛④——阳性

2 相关工作

2.1 基于医学对话的信息抽取相关研究

基于医疗文本的信息抽取任务发展已经相对成熟[1]，然而基于医学对话的信息抽取技术才刚刚起步，尤其在中文医学文本处理社区中，针对医学对话的研究还相对较少。Jeblee S 等[2]发布了一个系统来解析医学对话并提取药物、症状等实体，通过使用上下文判断实体的相关性，对医学对话的主要诊断进行分类，还提取对话中的主题信息并识别相关话语。Du N 等[3]从临床对话中抽取实体、属性及其关系，这个任务比一般情况下限制在几个相邻句子内推断实体之间的关系任务更复杂，因为需要

考虑更远距离的关系。Kannan A 等[4]采用半监督方式从医学对话中抽取症状信息、症状状态以及判断该症状患者是否正在经历。Patel D 等[5]使用弱监督技术从医学对话中自动提取用药方案，不仅可以提高召回率，而且帮助患者执行护理计划，减轻医生负担。Enarvi S 等[6]采用序列 - 序列模型从医学对话中自动生成医疗报告。

2.2 中文医学文本处理社区医学对话相关研究

对于中文医学对话也有相关工作来促进 NLP 社区发展。Lin X 等[7]构造了一个中文基于对话的症状诊断数据集供评估使用。Chen S 等[8]构造了一个大规模的中文和英文医疗对话数据集供研究使用，其中中文数据集来自“好大夫在线”网站（https://www.haodf.com/），英文数据集来自于 icliniq（https://www.icliniq.com/）和 healthcaremagic（https://www.healthcaremagic.com/）网站。Zhang Y 等[9]基于医学对话提出了一个医学信息提取器 MIE，能够提取提及的症状、手术、测试、其他信息及其相应状态，该数据集来自“春雨医生”（https://www.chunyuyisheng.com/）网站。第 20 届中国计算语言学大会（The Twentieth China National Conference on Computational Linguistics，CCL 2021）也发布了一个公开评测，即智能医疗对话诊疗公开评测，主要包括识别重要的医疗相关实体、对话意图、症状，生成医疗报告，判断疾病并给出相应建议。CHIP 2021 评测任务 1 基于医学对话的临床发现阴阳性判别任务对其中的疾病和症状等给出了更细粒度的阴阳性标注。

3 评测数据集

CHIP 2021 评测任务 1 的标注数据集全部来源于“春雨医生”的互联网在线问诊公开数据。标注数据由 1 名经验丰富的全职医学专家、6 名医学院学生以及 1 名算法专家全程参与，5 周时间内完成标注 10 000 段医疗对话，标注语料一致性约为 0.85。每条数据包括医生和患者的多轮对话，包括多个句子，每个句子可能包含多个医疗发现提及，也可能不包含医疗发现提及。每个实体提及需要被标注为“阳性”“阴性”“其他”或“不标注”4 个类别中的 1 个，见表 2。数据分为训练集、A 榜测试集和 B 榜测试集，数据大小分别为 6 000 条、2 000 条、1 999 条。根据数据集可以看出医疗对话长度偏长且标签分布极不均衡，见表 3。

表 2 不同类别的标注标准

类别	标注标准
阳性	已有症状疾/病等相关，医生诊断（包含多个诊断结论），以及假设未来可能发生的疾病等，如“如果不治疗的话，大概率会引起 A 疾病”，“A 疾病”标注为阳性
阴性	未患有的疾病症状相关
其他	未知的标注为其他，一般指用户没有回答、不知道或者回答不明确/模棱两可不好推断的情况
不标注	无实际意义的不标注，一般是医生的解释，说的是一般知识，和患者当前的状态条件独立不具有标注意义，及有些检查项带疾病名称而识别的疾病（乙肝五项/乙肝抗体中的“乙肝”），药品名中出现的疾病不标注

表 3 评测数据分布统计

数据集	数据量（条）	完整对话平均字符数（个）	单轮对话平均字符数（个）	完整对话最大字符数（个）	单轮对话最大字符数（个）	平均对话轮次（轮）	最大轮次（轮）	标签数量分布（阳性/阴性/其他/不标注）（条）
训练集	4 000	533.58	17.18	4 209	2 402	31.05	218	74 772/14 086/6 167/23 949
测试集 A	2 000	529.22	17.29	4 304	750	30.60	206	25 175/4 481/1 942/7 606
测试集 B	1 999	522.52	17.40	4 151	1 229	30.03	217	25 476/4 594/2 087/8 133

4 评估指标

CHIP 2021 评测任务 1 采用 Macro – $F1$ 作为最终评测指标。假设需要预测的 n（$n=4$）个类别标签分别为 $C_1, C_2, \ldots, C_n$，对于类别 C_k 的准确率和召回率计算方式如下：

$$\text{准确率 } P_k = \frac{\text{正确预测为类别 } C_k \text{ 类的样本个数}}{\text{预测为 } C_k \text{ 类的样本个数}} \quad (1)$$

$$\text{召回率 } P_k = \frac{\text{正确预测为类别 } C_k \text{ 类的样本个数}}{\text{真实的 } C_k \text{ 类的样本个数}} \quad (2)$$

则评测指标 Macro – $F1$ 计算方式如下：

$$\text{Macro} - F1 = \frac{1}{n}\sum_{k=1}^{n}\frac{2 * P_k * R_k}{P_k + R_k} \quad (3)$$

5 评测结果

5.1 整体情况

CHIP 2021 评测 1 评估分为 A 榜和 B 榜，主办方规定每支队伍每天最多提交 5 份结果，取 Macro – $F1$ 值最高的为该队伍最高结果，最终结果以 B 榜排名作为排名依据。A 榜评测共收到 81 支队伍的提交结果，B 榜共收到 15 支队伍的提交结果。根据 A、B 榜提交结果的整体情况可以看出，A、B 榜的最高分基本保持一致，说明数据的分布比较一致，见表 4。

表 4 A、B 榜结果整体情况（%）

榜单	最高分	最低分	平均分	中位数
A 榜	78.03	6.47	61.39	68.52
B 榜	78.10	24.29	67.23	70.43

5.2 评测方法

5.2.1 概述 对有效前 4 名参赛者的分数进行统计，前 4 名参赛者提交的模型均基于双向编码器表征（bidirectional encoder representations from transformers，BERT）[10]，可见 BERT 在中文自然语言处理任务中越来越受欢迎并且非常有效，见表 5。

表 5 前 4 名参与者分数

排名	单位	最高 Macro – $F1$（%）
第 1 名	祺鲸科技	78.10
第 2 名	卫宁健康科技集团股份有限公司	77.87
第 3 名	四川久远银海软件股份有限公司	76.30
第 4 名	厦门大学	75.40

5.2.2 方法 1 第 1 名参赛者（开源代码地址为 https://github.com/DataArk/CHIP2021 – Task1 – Top1）将该任务建模成细粒度的情感分析任务，在 4 种不同的中文 BERT 预训练语言模型上进行初始化，输入文本为当前角色与内容拼接，同时输入上下文信息，每个句子加入临床发现提及的标准化实体信息以及对话角色信息来增强模型的表示能力。这 4 种不同的中文预训练语言模型为 MedBERT[11]、MC – BERT[12]、MacBERT – Large[13]以及基于任务训练语料再训练掩码语言模型的 MacBERT – Large。训练时在每个预训练语言模型上采用 10 折交叉策略，得到 10 组模型，将 10 组模型进行投票，即每条数据将预测结果中出现标签最多的标签作为最终结果，如果次数相同则将每组模型中出现次数总和排前的作为预测标签。为了保证少样本标签“不标注”和“其他”能够预测出来，提出弱监督投票策略，即如果 MacBERT – Large 预测的 10 组结果中有 2 组及以上模型预测样本为“不标注”或“其他”，则该样本预测为“不标注”或“其他”。经过以上处理，该参赛者共得到 5 组结果，对这 5 组结果再次采用级联处理进行筛选修正，得到最终预测结果。由于真实数据标注可能存在一定噪声，该参赛者首先对数据进行清洗，在训练集训练并预测训练集，剔除预测结果与真实标签不一致的数据。最后在此基础上增加数据增强策略，以增强模型的泛化性和鲁棒性。

5.2.3 方法 2 第 2 名参赛者（开源代码地址为 https://github.com/winninghealth/chip2021）将该任务抽象化为对话信息抽取任务，识别以患者为主体、临床发现为客体的主客体之间关系。设计了角色感知的模型输入，包括显式输入和隐式输入。显式输入为每轮对话信息添加角色特殊符，隐式输入

则对每个输入字符加上角色编码嵌入。除此之外，该参赛者提出加入辅助序列来对每一个临床发现进行预测，具体做法是在多轮对话输入后加入患者角色的标志符以及要判别的临床发现。由于多轮对话中会出现临床发现提及相同但是阴阳性不同的情况，因此提出共享编码策略，即辅助序列中临床发现的编码共享该临床发现在多轮对话中相对应位置的编码，以此来避免歧义。由于预训练语言模型的输入有长度限制，在预处理阶段以临床发现提及为中心设计文本截断策略。为了缓解标签类别不平衡问题，提出交替归一化后处理，对低置信度预测结果进行修正。该参赛者对 6 个不同的中文预训练语言模型 CPT - BERT[14]、RoBERTa - Base[15]、RoBERTa - Large[15]、MacBERT - Base[13]、MacBERT - Large[13] 和 Chinese - BERT[16] 采用 5 折交叉验证进行微调测试并进行投票集成。

5.2.4 方法 3 第 3 名参赛者（开源代码地址为 https://github.com/mrgjbd/chip2021）为了充分利用预训练语言模型蕴含的知识，将该评测转化为掩码预测任务。其输入是多轮对话中的每轮对话拼接说话人角色信息，对于要预测的临床对话提及，输入最后拼接一个［E_ MASK］掩码进行预测，该掩码信息只可见对应的临床发现信息，其他均不可见。编码器部分则是采用可编码超长句子的 Longformer[17]，并在训练过程中加入对抗训练进行扰动，以增加模型的泛化性和鲁棒性。

5.2.5 方法 4 第 4 名参赛者（开源代码地址为 https://github.com/Yukie008/XMUchip2021）与第 3 名思路相似，将基于提示学习的医学对话临床发现阴阳性判别任务转化为掩码预测的完形填空任务，采用基于提示学习的方法，充分利用预训练语言模型的先验知识提升预测性能。对比基于预训练语言模型微调和基于预训练语言模型提示学习方法，发现提示学习可以很好地挖掘出预训练语言模型中潜在知识。此外该参赛者测试对比不同输入信息，例如拼接不同的上下文信息、拼接说话人角色信息等的性能，发现医学对话问诊数据具有事实倾向特征。

6 讨论

从参赛者方法看，值得注意的是，数据输入均采用特定的符号插入并区分医生和患者角色，说话者角色加入后能有效提高模型性能。在第 2 名参赛者方法中，显式地加入对话角色信息以及隐式地将对话角色信息编码嵌入加到每个字的嵌入中是一种有效的策略。不同对话角色通过不同的上下文信息输入对性能影响也很大，在第 1 名参赛者方法中，当是医生对话时，拼接下 3 轮患者对话作为上下文，而当是患者对话时，则不区分角色拼接下 3 轮对话。从参赛者提交的模型来看，参赛者均是基于医疗领域预训练模型 BERT 进行实验，甚至在任务数据上采用掩码预测任务再进行预训练，证明特定领域语料预训练的有效性。从第 3 名、第 4 名参赛者提交的结果看，采用提示模版的方法对任务进行建模较基于微调的方法有很大提升，主要是因为能够与预训练语言模型的预训练任务保持一致，充分利用预训练语言模型的潜在知识。从结果来看，每组参赛者的最终结果均是通过多个模型集成得到，集成带来了较大的提升。总体来说，前 4 名参赛者的结果相差并不明显，Macro - $F1$ 值均在 75% ~ 80% 之间。

7 结语

本文介绍 CHIP 2021 评测任务 1 的数据和结果。该任务是基于医学对话的临床发现阴阳性判别任务，评测 A、B 榜共收到近百份提交结果，其中最好的结果 Macro - $F1$ 值为 78.05%。该任务是中文医疗社区首次公开的基于医疗对话的临床发现阴阳性判别任务，后续数据集开放在 CBLUE（https://tianchi.aliyun.com/dataset/dataDetail?dataId=95414）排行榜中以推进医疗人工智能领域的发展。

参考文献

1 UZUNER Ö, SOUTH B R, SHEN S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical

text [J]. Journal of the American medical informatics association, 2011, 18 (5): 552 - 556.

2 JEBLEE S, KHAN KHATTAK F, CRAMPTON N, et al. Extracting relevant information from physician - patient dialogues for automated clinical note taking [C]. Hong Kong: Association for Computational Linguistics, 2019.

3 DU N, WANG M, TRAN L, et al. Learning to infer entities, properties and their relations from clinical conversations [C]. Hong Kong: Association for Computational Linguistics, 2019.

4 KANNAN A, CHEN K, JAUNZEIKARE D, et al. Semi - supervised learning for information extraction from dialogue [C]. Hyderabad: Interspeech, 2018.

5 PATEL D, KONAM S, PRABHAKAR S. Weakly supervised medication regimen extraction from medical conversations [C]. Online: Association for Computational Linguistics, 2020.

6 ENARVI S, AMOIA M, DEL - AGUA TEBA M, et al. Generating medical reports from patient - doctor conversations using sequence - to - sequence models [C]. Online: Association for Computational Linguistics, 2020.

7 LIN X, HE X, CHEN Q, et al. Enhancing dialogue symptom diagnosis with global attention and symptom graph [C]. Hong Kong: Association for Computational Linguistics, 2019.

8 CHEN S, JU Z, DONG X, et al. MedDialog: a large - scale medical dialogue dataset [C]. Online: The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

9 ZHANG Y, JIANG Z, ZHANG T, et al. MIE: a medical information extractor towards medical dialogues [C]. Online: Association for Computational Linguistics, 2020.

10 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, 2019.

11 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): 86.

12 ZHANG N, JIA Q, YIN K, et al. Conceptualized representation learning for Chinese biomedical text mining [EB/OL]. [2022 - 09 - 21]. https://www.researchgate.net/publication/343877180_Conceptualized_Representation_Learning_for_Chinese_Biomedical_Text_Mining.

13 CUI Y, CHE W, LIU T, et al. Revisiting pre - trained models for Chinese natural language processing [C]. Online: Association for Computational Linguistics, 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 - 09 - 21]. https://arxiv.org/pdf/2109.05729.pdf.

15 LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach [EB/OL]. [2022 - 09 - 21]. https://arxiv.org/pdf/1907.11692v1.pdf.

16 CUI Y, CHE W, LIU T, et al. Pre - training with whole word masking for Chinese BERT [EB/OL]. [2022 - 09 - 21]. https://arxiv.org/pdf/1906.08101v1.pdf.

17 BELTAGY I, PETERS M E, COHAN A. Longformer: the long - document transformer [EB/OL]. [2022 - 09 - 21]. https://arxiv.org/pdf/2004.05150v2.pdf.