# 基于医患对话的临床发现识别与阴阳性判别

林浩楠, 谭红叶, 冯慧敏

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231000084

# 基于医患对话的临床发现识别与阴阳性判别

林浩楠[1] 谭红叶[1,2] 冯慧敏[1]
1 山西大学计算机与信息技术学院 太原 030006
2 山西大学计算智能与中文信息处理教育部重点实验室 太原 030006
(202122407030@email.sxu.edu.cn)

**摘 要** 临床发现识别与阴阳性判别是智慧医疗领域的重要任务之一,旨在识别医患对话等医疗文本中的疾病与症状,并判断其阴阳性状态。该任务的现有研究主要不足有:(1)缺乏对医患对话语义信息、对话结构等特征的建模,导致模型准确率不高;(2)将该任务分为识别与判别两阶段进行,引起错误累积问题。针对以上不足,提出结合对话信息的统一生成模型,通过构建静态-动态融合图对医患对话语义、结构等信息建模,增强模型的对话理解能力;使用生成式语言模型将临床发现识别与阴阳性判别两个子任务统一为一个序列生成任务,以缓解错误累积问题,并且通过识别阴阳性指示词,辅助模型提高阴阳性判别准确率。在CHIP2021评测数据集CHIP-MDCFNPC上的实验结果表明:所提方法在F1指标上达到了71.83%,比基线模型平均提升了2.82%。

**关键词**:医患对话;临床发现识别;阴阳性判别;对话建模;统一生成模型
**中图分类号** TP391

# Clinical Findings Recognition and Yin & Yang Status Inference Based on Doctor-Patient Dialogue

LIN Haonan[1], TAN Hongye[1,2] and FENG Huimin[1]
1 School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
2 Key Laboratory of Ministry of Education Intelligence and Chinese Information Processing, Shanxi University, Taiyuan 030006, China

**Abstract** Clinical findings recognition and Yin & Yang status inference are import tasks in the field of intelligent healthcare. The goal is to identify clinical findings such as diseases and symptoms from doctor-patient dialogue record, then determine their Yin & Yang status. The main weakness of existing research is as follow: (1) Lack of modeling of semantic information and dialogue structure in doctor-patient dialogues, leading to low model accuracy. (2) Implementing it as a two-stage process, it will cause error accumulation. This paper proposes a unified generative method that incorporates dialogue information. It achieves this by constructing a static-dynamic fusion graph to model semantic and structural information in doctor-patient dialogues, enhancing the model's understanding of conversations. And utilizes a generative language model to unify clinical findings recognition and Yin & Yang status inference into a sequence generation task, mitigating the problem of error accumulation. Additionally, it improves the accuracy of Yin & Yang statue inference by identifying Yin & Yang statue indicator words. Experimental results on the CHIP2021 evaluation dataset CHIP-MDCFNPC show that the proposed method achieves an F1 score of 71.83%, which is 2.82% higher than the baseline model on average.

**Keywords** Doctor-Patient dialogue, Clinical findings recognition, Yin & Yang statue inference, Dialogue modeling, Unified generative model

## 1 引言

智慧医疗系统包括智能分诊、在线挂号、自动问诊和自助缴费等子系统,其中临床发现识别与阴阳性判别是自动问诊系统的核心任务之一。该任务的目标是识别医疗文本中的临床发现并判断阴阳性状态,它可以为自动问诊系统提供支持,提高医疗工作者的效率,缓解医疗资源紧张。

临床发现识别与阴阳性判别包含两个子任务:(1)临床发现识别,即识别出医疗文本中出现的疾病与症状;(2)阴阳性判别,即根据医疗文本信息推断临床发现的状态,患者有此病症为阳性,没有为阴性。表1列出了一些医患对话的示例,示例1中"发热"为症状,感冒"为疾病,均为临床发现,根据对话信息判断这两个临床发现的阴阳性状态分别为"阴性"和"阳性"。

基金项目:国家自然科学基金面上项目(62076155)
This work was supported by the General Program of the National Natural Science Foundation of China(62076155).
通信作者:谭红叶(hytan_2006@126.com)

表 1 基于医患对话的临床发现识别与阴阳性判别示例

Table 1 Example of clinical findings recognition and Yin & Yang status inference

| 序号 | 文本 | 临床发现 | 阴阳性状态 |
|---|---|---|---|
| 示例 1 | 医生:宝宝最近有感冒吗?有没有发热症状?<br>患者:有感冒但是没有发热。 | 症状:发热<br>疾病:感冒 | 发热:阴性<br>感冒:阳性 |
| 示例 2 | 医生:有咳嗽吗?<br>患者:有一点点。 | 症状:咳嗽 | 咳嗽:阳性 |
| 示例 3 | 医生:有咳痰吗?有没有腹部不适,腹痛等?<br>患者:有一点痰,宝宝也说不清肚子有什么不适。<br>医生:先服用感冒药治疗,观察后续情况。 | 症状:咳痰<br>症状:腹痛<br>疾病:感冒 | 咳痰:阳性<br>腹痛:其他<br>感冒:不标注 |

注:1)示例来自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评测任务。

2)其他:表示患者表述不清,无法判断。不标注:表示与患者当前状态无关。

当前,针对该任务的研究方法可以分为两类。第一类是基于两阶段的方法。例如,Du 等[1-2] 和 Lin 等[3] 在第一阶段使用序列标注模型进行命名实体识别得到临床发现,在第二阶段对临床发现进行分类得到不同的阴阳性状态。但是这种方法容易引起错误累积问题,同时忽略了不同临床发现之间的联系。第二类方法是生成式方法,该方法将临床发现识别与阴阳性判别两个子任务统一视为一个序列生成问题。例如,Finly 等[4] 和 Li 等[5] 把临床发现术语和阴阳性状态视为目标词汇,按顺序进行生成。这种方法还可以通过后续生成内容与先前内容的依赖关系,对不同临床发现之间的联系进行建模。

然而,无论是两阶段方法还是生成式方法,现有的研究大多将医疗对话视为单一的连续文本,忽略了对话的结构特征,缺乏对话语句间的语义、对话结构和角色转换等信息的建模。

本文提出结合对话信息的统一生成模型进行临床发现识别与阴阳性判别,主要思路如下:(1)使用静态-动态融合图建立对话语义、关键词共现和对话语句位置静态信息,并利用注意力机制建立对话的动态信息,增强模型的对话理解能力;(2)使用 Seq2Seq 生成式语言模型顺序生成临床发现术语及其阴阳性状态。此外,通过分析医疗文本发现,阴阳性状态常常可以通过识别一些指示词(如表 1 中的"有一点点""没有""有""说不清"等)来判断,因此本文引入阴阳性指示词识别,辅助模型进行阴阳性判别。

在 CHIP2021 评测数据集 CHIP-MDCFNPC 上的实验结果表明:本文所提方法在 F1 指标上达到了 71.83%,比基线模型平均提升了 2.82%。

## 2 相关工作

### 2.1 临床发现识别与阴阳性判别

临床发现识别与阴阳性判别是自动问诊系统的重要任务,可以减轻医生的工作强度,缓解医疗资源的紧缺,具有重要的研究价值。目前,研究者们主要通过两阶段方法和生成式方法进行研究。

针对两阶段方法,Du 等[1-2] 在第一阶段通过序列标注模型对临床发现进行识别,在第二阶段使用两个多层感知器分类器进行阴阳性状态判别;Lin 等[3] 通过 Bi-LSTM+CRF[6] 识别临床发现并构建症状关系图,之后通过分类器结合症状关系图对临床发现的阴阳性状态进行推断。Zhang 等[7] 使用基于信息抽取方法,使用结合对话交互信息的深度匹配模型抽取临床发现及其阴阳性状态。但是两阶段方法容易出现错误累积问题。

为了缓解上述问题,生成式方法逐渐进入视野。Finly 等[4] 提出了从自动语音识别转录到临床记录的生成方法,但效果较差;Li 等[5] 设计了基于相对位置编码的多粒度 Transformer[8] 模型生成临床发现和阴阳性状态,取得了不错的效果;Du 等[1] 也使用结合预训练的生成方法,取得了不弱于两阶段方法的结果。

### 2.2 生成式语言模型

生成式语言模型在自然语言处理领域占据重要地位,能够在各种应用场景中提供灵活多样、富有创造性的文本。其应用场景包括机器翻译、摘要生成和问答等,具有广阔的应用前景。通过生成式模型完成任务已成为自然语言处理领域的一种重要范式。

当前主流的生成模型分为两类:Seq2Seq 模型和解码器模型。Seq2Seq 模型由编码器-解码器构成,编码器负责将输入序列编码为一个固定长度的向量表示,而解码器则将该向量解码为目标序列。早期的 Seq2Seq 模型通常使用循环神经网络(RNN)[9]、长短期记忆网络(LSTM)[10] 等来实现。目前应用广泛的是基于预训练 Transformer 的 Seq2Seq 模型,其在大规模无标注数据上进行预训练,有强大的自然语言理解和生成能力。典型模型有 T5[11] 和 BART[12] 等,T5 采用自注意力的提示生成策略,以整个序列为单位进行生成;BART 采用自回归方式逐字符生成文本,在自然语言生成任务上有很好的表现。

解码器模型,又称仅解码器模型(Decoder-Only),通常不包含编码器,由 Transformer 解码器组成,采用自注意力机制理解并利用上下文信息,通过前馈神经网络生成内容丰富、上下文一致的文本。当前应用最广泛的仅解码器生成模型是 GPT 系列[13-15],其同样采用自回归的生成策略。

由于任务特点,模型需要很好地理解输入信息,并在输出中利用这些信息,因此本文使用 Seq2Seq 模型 BART 作为结合对话信息的统一生成模型的基干。

### 2.3 对话建模

对话是自然语言的基本形式,具有动态交互的性质,信息流分散在不同对话参与者的多个语句中。将对话文本作为单一的连续文本处理会忽略对话的结构特征和对话语句间信息的交互与转移,极大地影响模型在对话任务上的表现。为了解决这个问题,当前研究主要通过外部对话解析工具构建启发式静态图建模对话信息,主要目标包括:对话语义解析[16-17]、对话主题建模[18-19]、对话状态跟踪[20] 和对话行为建模[16,21]。静态图在一定程度上捕获了对话语句间信息的交互和转移,并且在大量实验中证明了有效性,但是存在以下问题:(1)外部对话解析工具可能无法提供准确的解析结果,在数据分布极大地变化时,其稳定性和泛化能力会受到考验;

(2)静态图的构建独立于图表示学习阶段，其固定不变的特性使其不能动态地关注对话的不同信息。

在此基础上，Gao 等[22]提出基于静态-动态融合图的对话建模方法，在原有静态图的基础上通过多头注意力机制构建动态图，捕捉启发式静态图无法关注到的信息，进一步丰富对话表示。

本文受 Gao 等[22]对话建模方法的启发，将其与生成式语言模型 BART 融合，形成了结合对话信息的统一生成模型。该模型利用静态-动态融合图对医患对话进行建模，获取对话的结构特征和语义信息，并通过生成式语言模型将临床发现识别与阴阳性判别任务包含的两个子任务统一为一个序列生成任务，在缓解错误累积问题的同时建立不同临床发现间的联系。此外，在引入静态-动态融合图时，我们发现医患对话仅包含两个角色，没有复杂的角色交互，故在建模时去除了对话角色关系静态图。同时，在生成阶段引入了阴阳性指示词识别，通过生成临床发现对应的指示词，辅助模型进行阴阳性判别，提升准确率。

## 3 方法

### 3.1 任务定义

临床发现识别与阴阳性判别任务可以形式化定义为：给定一个医患对话文本 $X=\{x_1,\cdots,x_{L_d}\}$，$x_i=[w_{i,1},\cdots,w_{i,L_u}]$，阴阳性类别 $C=\{c_1,\cdots,c_L\}$，求目标序列 $Y=[e_1,s_1,\cdots,e_m,s_m]$。其中 $x_i$ 表示对话中的第 $i$ 个语句，共有 $L_d$ 个语句；$w_{i,j}$ 表示语句 $x_i$ 的第 $j$ 个词，$L_u$ 表示语句的长度；$c_i$ 表示第 $i$ 个类别，$L$ 表示阴阳性类别数目。$Y$ 中的 $e_i$ 和 $s_i$ 分别表示第 $i$ 个临床发现及其阴阳性状态，共有 $m$ 个。候选词表来自于医疗对话和阴阳性类别，即 $Y\in X\cup C$。该任务的表示如式(1)所示：

$$P(Y|X)=\prod_{t=1}^{m}P(y_t|X,Y_{<t}) \tag{1}$$

### 3.2 结合对话信息的统一生成模型

本文采用的结合对话信息的统一生成模型如图 1 所示，分为 3 个模块：对话编码模块、静态-动态图融合模块、解码生成模块。

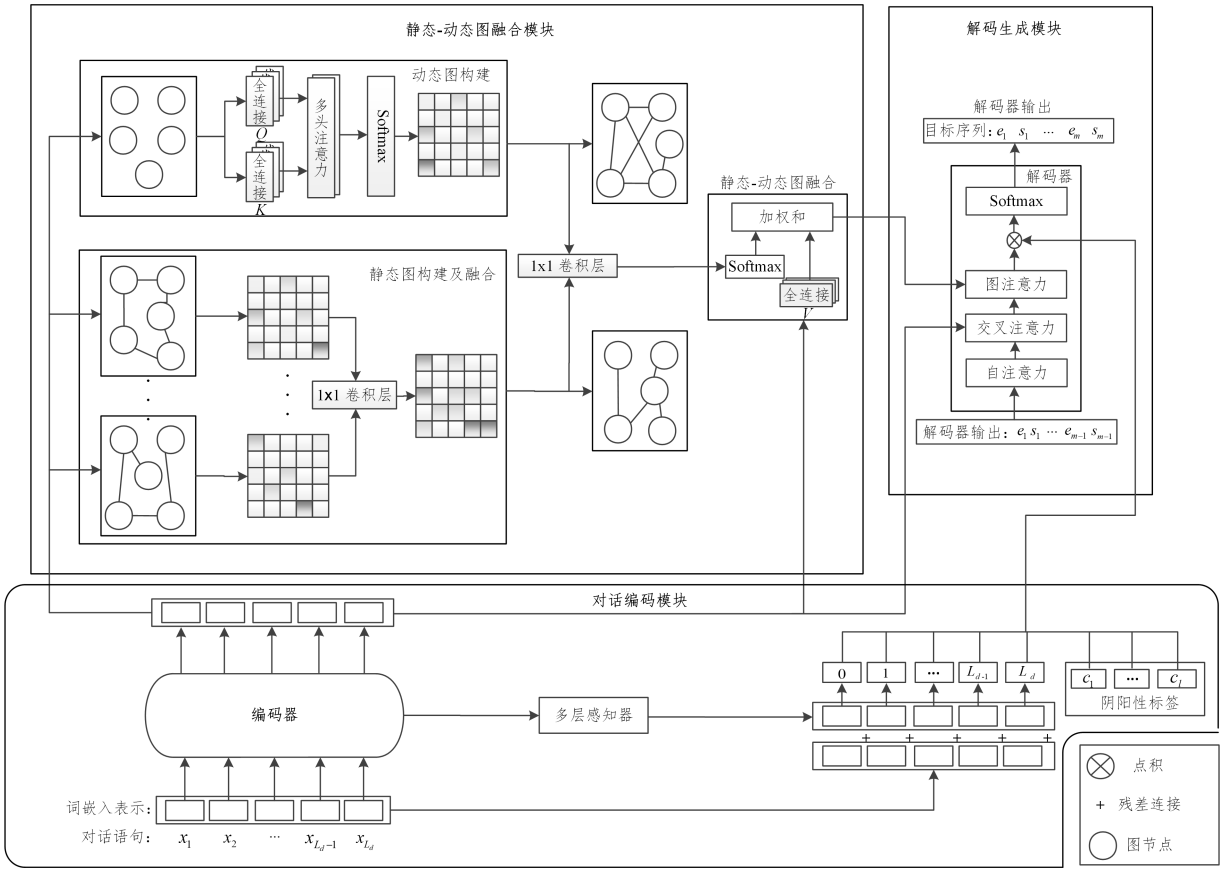


图 1 结合对话信息的统一生成模型

Fig. 1 Unified generative model incorporates with dialogue information

#### 3.2.1 对话编码模块

本文使用预训练 BART 编码器对每个对话语句按照式(2)进行独立编码：

$$\boldsymbol{u}_i=BARTEncoder([CLS],w_{i,1},\cdots,w_{i,L_u}) \tag{2}$$

其中，$\boldsymbol{u}_i$ 为医患对话第 $i$ 句话的向量表示，$L_u$ 表示语句长度，$\boldsymbol{u}_i\in\mathbb{R}^{L_u\times d}$，$d$ 为隐藏层维度。最终得到全部对话语句的向量表示 $\boldsymbol{U}=\{u_1,\cdots,u_{L_d}\}$。

首先，将编码器输出的对话向量表示 $\boldsymbol{U}$ 输入多层感知器(Multilayer Perceptron，MLP)，得到 $\hat{\boldsymbol{U}}$，计算式如式(3)所示：

$$\hat{\boldsymbol{U}}=MLP(\boldsymbol{U}) \tag{3}$$

其次，将 $\hat{\boldsymbol{U}}$ 与对话 $\boldsymbol{U}$ 的词嵌入表示 $\boldsymbol{E}$ 进行残差连接得到 $\bar{\boldsymbol{U}}$，$\alpha$ 为残差连接的参数。如式(4)、式(5)所示：

$$\boldsymbol{E}=TokenEmbed(X) \tag{4}$$

$$\bar{\boldsymbol{U}}=\alpha\hat{\boldsymbol{U}}+(1-\alpha)\boldsymbol{E} \tag{5}$$

最后，将阴阳性类别的词嵌入表示 $\boldsymbol{C}^e$ 与 $\bar{\boldsymbol{U}}$ 拼接得到 $\tilde{\boldsymbol{U}}$，如式(6)、式(7)所示：

$$\boldsymbol{C}^e=TokenEmbed(C) \tag{6}$$

$$\widetilde{\boldsymbol{U}}=[\bar{\boldsymbol{U}};\boldsymbol{C}^e] \tag{7}$$

其中，$\hat{\boldsymbol{U}},\bar{\boldsymbol{U}},\widetilde{\boldsymbol{U}},\boldsymbol{E}\in\mathbb{R}^{L_d\times d},\boldsymbol{C}^e\in\mathbb{R}^{l\times d}$，$\alpha$ 为参数。

3.2.2 静态动态图融合模块

(1)静态图构建与融合

本文使用了 3 种启发式静态对话图：语句关系图、关键词共现图、对话语句位置图。

①语句关系图。主要建模对话语句之间(包括非相邻语句)的关系。本文主要关注解释、问答、叙述 3 种关系。其中，解释是指医生解释患者的提问或医学知识；问答是指患者回答医生的提问；叙述是指患者自述病情。3 种关系的示例如表 2 所示。使用语句解析工具[23]构建基于依赖的对话结构信息，获得语句关系后使用嵌入矩阵将这些离散的关系映射为向量表示。具体操作如式(8)所示：

$$\boldsymbol{g}_d^s(i,j)=\varepsilon_d(DiscoParse(x_i,x_j)) \tag{8}$$

其中，$\boldsymbol{\varepsilon}_d\in\mathbb{R}^{3,1}$ 表示嵌入矩阵，$DiscoParse(\cdot)$ 表示语句解析工具输出的语句关系，$x_i$ 和 $x_j$ 分别表示对话中的第 $i$ 与第 $j$ 个句子。

表 2 语句关系示例

Table 2 Examples of discourse relations

| 序号 | 文本 | 语句关系 |
|---|---|---|
| 示例 1 | 患者：为什么最近心跳很快，像刚做过剧烈运动？<br>医生：静息心率过快可能是由窦性心动过速或者心肌炎造成的，具体情况需要进行检查才能确定。 | 解释 |
| 示例 2 | 医生：您好，请问肿胀持续多长时间了？<br>患者：有一星期左右了。 | 问答 |
| 示例 3 | 患者：双侧腰部疼痛，有时候会引起腿疼。<br>患者：检查结果双侧竖脊肌多裂肌萎缩，请问这种有办法治疗吗？ | 叙述 |

②关键词共现图。主要建模对话语句间的关键词共现关系，当两句话中包含相同的关键词时，它们可能关注的是同一主题，并且在语义上具有相关性。本文定义关键词为临床发现，统计两个语句出现同一个临床发现的次数，使用嵌入矩阵将共现次数映射为向量表示，具体操作如式(9)所示：

$$\boldsymbol{g}_k^s(i,j)=\varepsilon_k(KeyCo\text{-}ooc(x_i,x_j)) \tag{9}$$

其中，$x_i,x_j$ 分别表示对话中的第 $i$ 与第 $j$ 个句子，$KeyCo\text{-}occ(\cdot)$ 表示计算 $x_i$ 和 $x_j$ 出现同一个关键词的次数；$\boldsymbol{\varepsilon}_k\in\mathbb{R}^{N_k,D}$ 表示嵌入矩阵，$N_k$ 表示关键词共现的最大次数，$D$ 表示隐藏层的维度。

③对话语句位置图。主要建模对话语句的位置信息，将对话之间的相对距离作为对话语句位置图 $g_p^s$ 的边。使用嵌入矩阵将离散的距离映射到向量空间，如式(10)所示：

$$\boldsymbol{g}_p^s(i,j)=\boldsymbol{\varepsilon}_p(j-i) \tag{10}$$

其中，$g_p^s$ 为对话语句位置信息图的邻接矩阵，值为对话语句 $i$ 和 $j$ 的相对距离，$\boldsymbol{\varepsilon}_p\in\mathbb{R}^{L_d,1}$ 表示嵌入矩阵。

在获得上述 3 个静态图后进行融合，将每个图的邻接矩阵视为一个通道，使用 1×1 的卷积层将这些邻接矩阵整合为一个融合了 3 种对话信息的表示，如式(11)所示：

$$\boldsymbol{g}^s=Conv(\boldsymbol{g}_d^s\oplus\boldsymbol{g}_k^s\oplus\boldsymbol{g}_p^s) \tag{11}$$

其中，$\oplus$ 表示矩阵拼接，$\boldsymbol{g}^s\in\mathbb{R}^{L_d\times L_d}$ 是融合了对话信息的静态表示。

(2)动态图构建

动态图不使用任何预先计算或启发式的方法构建图结构，完全使用注意力机制捕捉基于深度向量表示的语义关系。将对话向量表示 $\boldsymbol{U}=\{u_1,\cdots,u_{L_d}\}$ 映射到两个不同的向量空间，计算它们之间的关系矩阵 $\boldsymbol{A}$，如式(12)所示：

$$\boldsymbol{Q}=\boldsymbol{U}\boldsymbol{W}_Q,\boldsymbol{K}=\boldsymbol{U}\boldsymbol{W}_K,\boldsymbol{A}=\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}} \tag{12}$$

其中，$\boldsymbol{W}_Q$ 和 $\boldsymbol{W}_K$ 为可训练的参数，$d_K$ 为平滑因子。使用关系矩阵 $\boldsymbol{A}$ 作为对话动态图的邻接矩阵，该图通过注意力机制构建，具有可训练的参数，可以捕捉启发式静态图无法覆盖的语句间特定的关系，进一步丰富对话信息的表示。

(3)静态-动态图融合

为了实现静态图与动态图的融合，需要将动态图的关系矩阵 $\boldsymbol{A}$ 与静态图的邻接矩阵 $\boldsymbol{g}^s$ 合并为一个统一图 $\boldsymbol{g}^u$。与静态图融合方法相同，使用 1×1 的卷积层，将 $\boldsymbol{A}$ 和 $\boldsymbol{g}^p$ 视为两个通道进行合并，如式(13)所示：

$$\boldsymbol{g}^u=Conv(\boldsymbol{A}\oplus\boldsymbol{g}^s) \tag{13}$$

其中，$\boldsymbol{g}^u\in\mathbb{R}^{L_d\times L_d}$。

为了将静态-动态融合图整合到最终的对话语句表示中，本文模型采用自注意力层，如图 1 所示。首先使用与式(12)相同的多头注意力机制将对话语句表示映射到多个向量空间中，之后使用统一图 $g^u$ 作为注意力权重进行加权和(Weighted Sum)运算，如式(14)、式(15)所示：

$$\boldsymbol{V}=\boldsymbol{U}\boldsymbol{W}_V \tag{14}$$

$$\{\boldsymbol{g}_1,\cdots,\boldsymbol{g}_{L_d}\}=\mathrm{softmax}(\boldsymbol{g}^u)\boldsymbol{V} \tag{15}$$

其中，$\boldsymbol{W}_v$ 为可训练的参数，$\boldsymbol{g}_i$ 为对话中第 $i$ 个融合了静态-动态融合图信息的语句向量表示。

3.2.3 解码生成模块

解码生成模块的主要作用是结合对话编码信息和对话建模信息生成临床发现实体及其阴阳性状态。具体实现如下：

解码生成模块主要由自注意力层、交叉注意力层和图注意力层构成。首先将目标序列 $Y$ 前 $t-1$ 个时间段的输出 $Y_{<t}$ 输入自注意力层，得到自注意力层的输出 $\boldsymbol{p}^s$，然后将 $\boldsymbol{p}^s$ 与对话表示 $U$ 输入交叉注意力层，进行交叉注意力计算，如式(16)、式(17)所示：

$$\boldsymbol{p}^s=self\text{-}attention(Y_{<t}) \tag{16}$$

$$\boldsymbol{p}^q=MHAtt(\boldsymbol{p}^s,\boldsymbol{U}) \tag{17}$$

其中，$MHAtt$ 表示多头注意力层，与原始的 BART 解码器相同。

在交叉注意力层后应用图注意力层，根据每个解码步骤从不断更新的图中获取有用的对话信息，过程如式(18)所示：

$$\boldsymbol{p}^g=MHAtt(\boldsymbol{p}^q,\{\boldsymbol{g}_1,\cdots,\boldsymbol{g}_{L_d}\}) \tag{18}$$

其中，$\boldsymbol{p}^g$ 为图注意力层输出的隐状态表示。

最后，利用图注意力层的输出 $\boldsymbol{p}^g$ 和对话编码模块得到的 $\widetilde{\boldsymbol{U}}^e$ 做点积(Dot-product)，求 Softmax 得到 $y_t$ 在候选词表中的概率分布，具体计算如式(19)所示：

$$\boldsymbol{P}_t=\mathrm{Softmax}(\widetilde{\boldsymbol{U}}^e\boldsymbol{p}^g) \tag{19}$$

其中，$\boldsymbol{P}_t\in\mathbb{R}^{(L_d+l)}$，$\boldsymbol{P}_t$ 为在所有候选词表上的概率分布。

模型在训练阶段使用教师强制策略，并采用负对数似然进行优化；在生成阶段使用集束搜索以自回归的方式生成目标文本。

### 3.3 候选词表

模型在生成过程中按顺序生成临床发现、阴阳性指示词和阴阳性状态，候选词表来自于医患对话文本和阴阳性状态。在生成临床发现后，模型在临床发现附近的医疗文本中寻找

并生成阴阳性指示词,它是可以指示临床发现状态的关键性词或短语,主要包括以下 4 类:(1)时间词(如"现在""之前""这几天"等);(2)状态词(如"有""没有""考虑是"和"不是"等);(3)与治疗相关的词(如"有利于""恢复"等);(4)药物与检查(如"xx 药""xx 检查"等)。前两项作为阴性、阳性和其他三类的指示词,后两项作为不标注的指示词。指示词起到提示作用,与临床发现一起指导模型生成阴阳性状态。最终输出的结果中隐去阴阳性指示词。

### 3.4 损失函数

本文使用交叉熵(Cross-Entropy)作为损失函数训练模型,交叉熵是自然语言生成任务中最常用的损失函数之一,用于衡量生成的文本分布与真实文本分布之间的差异,通过最小化损失函数鼓励模型生成更接近目标的序列,如式(20)所示:

$$Loss=-\sum_{t=1}^{T}\boldsymbol{y}_t\log(\hat{\boldsymbol{y}}_t) \tag{20}$$

其中,$\boldsymbol{y}_t$ 和 $\hat{\boldsymbol{y}}_t$ 分别表示 $t$ 时刻的真实文本和模型生成的文本。

## 4 实验

### 4.1 数据集

本文使用的数据集为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评测任务数据集(CHIP-MDCFNPC),表 3 列出了数据集的统计信息。

阴阳性状态分为 4 类:阴性、阳性、其他和不标注。阳性:患者有此疾病或症状;阴性:患者没有此疾病或症状;其他:用户没有回答或者回答不明确,不好判断;不标注:医生解释医学知识或带有临床发现名称的药和检查项目,与病人当前状态无关。实例如表 3 所列。

表 3 CHIP-MDCFNPC 数据集概况

Table 3 Overview of CHIP-MDCFNPC dataset

| 子集 | 对话数 | 临床发现数量 |
|---|---|---|
| 训练集 | 6000 | 118976 |
| 验证集 | 1000 | 22154 |
| 测试集 | 2000 | 39204 |

### 4.2 评价指标

本文采用 F1 指标评价模型在临床发现识别和阴阳性判别任务中的表现,识别与判别结果都正确才视为正确,计算过程如式(21)—式(23)所示:

$$P=\frac{\text{正确预测的样本数}}{\text{预测出的样本数}} \tag{21}$$

$$R=\frac{\text{正确预测的样本数}}{\text{全部样本数}} \tag{22}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{23}$$

其中,$P$ 和 $R$ 表示模型在临床发现识别与阴阳性判别任务上的准确率和召回率。

同时,本文还采用 Macro-F1 指标评价模型在阴阳性判别任务上的表现,用以验证对话建模和阴阳性指示词的有效性,计算过程如式(24)—式(26)所示:

$$P_i=\frac{\text{正确预测为 } C_i \text{ 类的样本数}}{\text{预测为 } C_i \text{ 类的样本数}} \tag{24}$$

$$R_i=\frac{\text{正确预测为 } C_i \text{ 类的样本数}}{\text{真实 } C_i \text{ 类的样本数}} \tag{25}$$

$$Macro\text{-}F1=\frac{1}{N}\sum_{i=0}^{n}\frac{2\times P_i\times R_i}{P_i+R_i} \tag{26}$$

其中,$P_i$ 和 $R_i$ 分别为第 $i$ 个阴阳性类别的准确率和召回率,$N$ 表示类别总数。

### 4.3 基线模型

为了验证本文方法的有效性,将结合对话信息的统一生成模型与两阶段模型 K-BERT、SAT、FLAT-SAT、BERT-SAT、MIE 和 KTGF 以及生成式模型 ChatGPT 进行对比。模型介绍如下:

K-BERT[24]:结合知识图谱和 BERT[25] 预训练语言模型,通过融合语义表示和知识图谱的额外信息,提升了模型在医疗、金融等特定领域的表现,可完成命名实体识别和分类任务。

SAT[1]:基于两阶段的方法,第一阶段由序列标注模型(Bi-LSTM + CRF[26])使用 BIO 标签标注临床发现;第二阶段使用多层感知器分类器进行阴阳性判别。通过多任务联合学习的方式获得更好的效果。

FLAT-SAT:与 SAT 模型类似,在两阶段方法中采用更为先进的序列标注模型 FLAT[27] 进行临床发现识别,然后通过分类器进行阴阳性判别。

BERT-SAT:使用 BERT 作为编码器的两阶段模型,增强模型的自然语言理解能力,与 SAT 类似,第一阶段做序列标注识别临床发现,第二阶段利用分类器实现阴阳性判别。

MIE[7]:基于深度匹配方法的多标签分类方法,可以考虑对话回合的交互,在一定程度上建立对话信息。

KTGF[28]:基于知识增强的两阶段生成式医学信息抽取模型,在第一阶段利用 T5 生成仅包含临床发现的序列,并建模它们之间的关系;第二阶段设计提示模板,利用模型的先验知识更好地完成阴阳性判别任务。

ChatGPT:基于 GPT-3.5 架构的大语言模型,由于其具有强大的自然语言理解和生成能力,同时还能够理解自然语言指令,可以广泛应用于多种自然语言处理任务,该模型还通过了医学考试[29]。

### 4.4 实验细节

阴阳性指示词标注:(1)人工标注,通过 doccano 数据标注平台标注 500 组对话全部临床发现的阴阳性指示词;(2)自动标注,使用人工标注的数据对 UIE 通用信息抽取框架[30] 进行微调,然后对训练集进行自动标注;(3)数据检查及修改,打乱标注后的训练集,抽取 500 组对话进行人工核查,结果显示 80%以上的阴阳性指示词都能反映出当前临床发现的阴阳性状态,对标注错误进行修改后形成实验数据集。

ChatGPT 调用:通过调用 GPT-3.5 的 API,采用 In-Context Learning 的方式,激活思维链。具体步骤如下:(1)给出若干医疗文本,提示 ChatGPT 症状与疾病都属于临床发现;(2)将医疗对话文本、临床发现及其对应的阴阳性输入模型,让模型结合对话内容关注到不同临床发现的阴阳性状态;(3)将完整的示例输入模型并使用测试集进行测试。

实验设置:实验使用 Pytorch 深度学习框架,在 Ubuntu 系统上采用 NVIDIA A100 进行模型的训练和调试。本文使用的 BART-large-Chinese 生成式语言模型作为基础,输出词向量维度为 1024。学习率为 $5\times10^{-5}$,Batch Size 设置为 32,epoch 为 50。

### 4.5 实验结果与分析

表4列出了基线模型和本文方法在测试集上的实验结果，结合对话信息的统一生成模型在F1指标上取得了71.83%，在Macro-F1指标上取得了77.92%，对比两阶段模型在两个指标上平均分别提升了2.87%和3.27%，对比ChatGPT在两个指标上也提升了2.5%和1.96%。相比没有进行对话建模的基线，实验结果证明了本文方法在对话理解方面的优势，阴阳性判别准确率有不错的提升；同时，生成式方法有效缓解了错误累积问题，并且其通过识别阴阳性指示词，辅助模型进行阴阳性判别。

从实验结果可以看出，ChatGPT虽然是目前最先进的大语言模型，具有极强的对话理解和生成能力，但是由于只能通过API接口调用，利用其In-Context-Learning能力适应任务，无法进行全量参数微调，因此在特定领域任务的表现会略低于传统微调模型，但是依旧强于一些基于两阶段方法的模型。

表4 对比实验结果

Table 4 Results of comparative experiments

(%)

| 模型 | P | R | F1 | Macro-F1 |
|---|---|---|---|---|
| K-BERT | 68.01 | 68.84 | 68.42 | 73.76 |
| SAT | 69.15 | 68.43 | 68.79 | 74.46 |
| FLAT-SAT | 70.22 | 68.68 | 69.44 | 74.73 |
| BERT-SAT | 68.87 | 69.55 | 68.21 | 74.34 |
| MIE | 68.98 | 68.08 | 68.53 | 74.41 |
| KTGF | 69.57 | 71.19 | 70.37 | 76.21 |
| ChatGPT | 68.13 | 70.57 | 69.33 | 75.96 |
| Ours | **72.75** | **70.93** | **71.83** | **77.92** |

### 4.6 消融实验

为了进一步验证本文方法的有效性，我们对模型进行了消融实验，形成了4种基线，实验结果如表5所列。

表5 消融实验结果

Table 5 Results of ablation experiment

(%)

| 模型 | P | R | F1 | Macro-F1 |
|---|---|---|---|---|
| w/o 静态图 | 71.87 | 70.42 | 71.14 | 77.02 |
| w/o 动态图 | 72.08 | 70.50 | 71.28 | 77.08 |
| w/o 静态-动态融合图 | 70.31 | 69.14 | 69.72 | 75.43 |
| w/o 阴阳性指示词识别 | 71.58 | 69.73 | 70.64 | 76.81 |
| Ours | **72.75** | **70.93** | **71.83** | **77.92** |

可以看出，所有消融实验的结果均不如本文方法。动态图与静态图信息互补，共同进行对话建模，加入静态-动态融合图后模型的表现有较大提升。去除阴阳性指示词识别的消融实验在两个指标上均有一定程度的降低，证明了阴阳性指示词识别对模型进行阴阳性判别起到了辅助作用，提升了模型的表现。

为了验证静态图对模型的贡献，本文对每种静态图进行消融实验，结果如表6所列。可以发现，对话语句位置图的贡献最大，表明对话语句位置可以帮助模型理解对话结构。虽然关键词共现图可以直观地反映对话内容的相关性，但是只为最终结果贡献了0.58%的F1分数。此外，表6中的模型均优于仅使用动态图的模型(w/o 静态图)，证明了使用启发式静态图可以有效弥补模型中对话先验知识的空缺，直观地建模对话信息。

表6 针对不同静态图的消融实验结果

Table 6 Results of ablation experiments for different static graphs

(%)

| 模型 | P | R | F1 | Macro-F1 |
|---|---|---|---|---|
| w/o 语句关系图 | 72.24 | 70.59 | 71.41 | 77.28 |
| w/o 关键词共现图 | 72.29 | 70.71 | 71.49 | 77.34 |
| w/o 对话语句位置图 | 72.13 | 70.62 | 71.37 | 77.12 |
| Ours | **72.75** | **70.93** | **71.83** | **77.92** |

### 4.7 对话建模方法对比实验

为了进一步证明静态-动态融合图建模对话信息的有效性，本文将两种不同对话建模方法与统一生成模型结合，建模方法介绍如下：

SSAnet[31]：设计了一个异构的语义槽图，用于捕捉对话中的关键信息及其依赖关系，并利用掩码交叉注意力机制关注这些信息，增强槽特征以建立对话语句之间的关系。

Coref-Attn[32]：通过图神经网络(Graph Neural Network)建立角色类和事件之间的共指关系图，用于建模对话信息。

实验结果如表7所列，静态-动态融合图建模方法获得了更好的性能，表明本文所用方法能够更细致地融入对话结构和信息等先验知识，并且通过多头注意力机制动态捕捉对话的深层语义关系，有助于模型更好地理解对话的上下文信息。

表7 针对不同对话建模方法的实验结果

Table 7 Results of different dialogue modeling approaches

(%)

| 方法 | P | R | F1 | Macro-F1 |
|---|---|---|---|---|
| SSAnet | 71.93 | 70.15 | 71.03 | 77.08 |
| Coref-Attn | 72.09 | 70.50 | 71.29 | 77.15 |
| Ours | **72.75** | **70.93** | **71.83** | **77.92** |

### 4.8 错误分析

针对临床发现识别任务，从表8的错误样例分析中可以看出，模型的预测出现了一些错误，样例1包含“甲减”和“甲亢”两个临床发现，但模型结合语境错误地将其识别成“甲减甲亢”一个临床发现；在样例2中包含一个复杂临床发现“进行性肌阵挛癫痫”，模型仅识别出“癫痫”。

表8 临床发现识别错误样例分析

Table 8 Error analysis of clinical findings recognition

| 序号 | 医疗文本 | 临床发现 | 模型识别结果 |
|---|---|---|---|
| 样例1 | 医生：检查结果似乎没有甲减甲亢的症状。 | 疾病：甲减<br>疾病：甲亢 | 疾病：甲减甲亢 |
| 样例2 | 医生：考虑是进行性肌阵挛癫痫。 | 疾病：进行性肌阵挛癫痫 | 疾病：癫痫 |

其中主要错误原因在于模型缺少准确识别临床发现的医学知识，导致模型只能根据学习到的医疗文本规律进行识别，遇到复杂临床发现时仅能识别出常见部分。

**结束语** 本文提出了结合对话信息的统一生成模型来处理临床发现识别与阴阳性判别任务，通过静态-动态融合图建模对话中的对话语义、关键词共现、对话语句位置等静态信息，并使用注意力机制动态捕捉对话信息，增强模型的对话理解能力，同时，将两阶段任务建模为生成式任务，使用Seq2Seq模型顺序生成临床发现术语及其阴阳性状态，缓解错误累积问题并关注不同临床发现之间的联系，结合阴阳性指示词识别，辅助模型进行阴阳性判别，提高了判别的准确率。实验结果表明，本文方法在临床发现识别与阴阳性判别任务上取得了不错的性能提升。

下一步工作中，我们将尝试引入医疗知识图谱等外部知识库，减少因缺乏领域知识带来的模型性能损失，增强在临床发现识别上的表现。

## 参 考 文 献


[1] DU N，CHEN K，KANNAN A，et al. Extracting symptoms and their status from clinical conversations[J]. arXiv：1906.02239，2019.

[2] DU N，WANG M，TRAN L，et al. Learning to infer entities，properties and their relations from clinical conversations[J]. arXiv：1908.11536，2019.

[3] LIN X，HE X，CHEN Q，et al. Enhancing dialogue symptom diagnosis with global attention and symptom graph[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019：5033-5042.

[4] FINLEY G，SALLOUM W，SADOUGHI N，et al. From dictations to clinical reports using machine translation[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，Volume 3(Industry Papers). 2018：121-128.

[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] HUANG Z，XU W，YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv：1508.01991，2015.

[7] ZHANG Y，JIANG Z，ZHANG T，et al. MIE：A medical information extractor towards medical dialogues[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020：6460-6469.

[8] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J]. arXiv：1706.03762，2017.

[9] ELMAN J L. Finding structure in time[J]. Cognitive science，1990，14(2)：179-211.

[10] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural computation，1997，9(8)：1735-1780.

[11] RAFFEL C，SHAZEER N，ROBERTS A，et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. J. Mach. Learn. Res.，2020，21(140)：1-67.

[12] LEWIS M，LIU Y，GOYAL N，et al. BART：Denoising Sequence-to-Sequence Pre-training for Natural Language Generation，Translation，and Comprehension[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020：7871-7880.

[13] RADFORD A，NARASIMHAN K，SALIMANS T，et al. Improving language understanding by generative pre-training[J/OL]. https：//openai.com/research/language-unsupervised.

[14] RADFORD A，WU J，CHILD R，et al. Language models are unsupervised multitask learners[J]. OpenAI blog，2019，1(8)：9.

[15] BROWN T，MANN B，RYDER N，et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems，2020，33：1877-1901.

[16] CHEN J，YANG D. Structure-aware abstractive conversation summarization via discourse and action graphs[J]. arXiv：2104.08400，2021.

[17] FENG X，FENG X，QIN B，et al. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization[J]. arXiv：2012.03502，2020.

[18] CHEN J，YANG D. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization[J]. arXiv：2010.01672，2020.

[19] ZHAO L，XU W，GUO J. Improving abstractive dialogue summarization with graph structures and topic words[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020：437-449.

[20] ZHAO L，ZHENG F，HE K，et al. Todsum：Task-oriented dialogue summarization with state tracking[J]. arXiv：2110.12680，2021.

[21] GOO C W，CHEN Y N. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts[C]// 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE，2018：735-742.

[22] GAO S，CHENG X，LI M，et al. Dialogue Summarization with Static-Dynamic Structure Fusion Graph[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics(Volume 1：Long Papers). 2023：13858-13873.

[23] SHI Z，HUANG M. A deep sequential model for discourse parsing on multi-party dialogues[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019：7007-7014.

[24] LIU W，ZHOU P，ZHAO Z，et al. K-bert：Enabling language representation with knowledge graph[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020：2901-2908.

[25] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv：1810.04805，2018.

[26] LAFFERTY J，MCCALLUM A，PEREIRA F C N. Conditional random fields：Probabilistic models for segmenting and labeling sequence data[C]// ICMI. 2001.

[27] LI X，YAN H，QIU X，et al. FLAT：Chinese NER using flat-lattice transformer[J]. arXiv：2004.11795，2020.

[28] HU Z，NI Z，SHI J，et al. A Knowledge-enhanced Two-stage Generative Framework for Medical Dialogue Information Extraction[J]. arXiv：2307.16200，2023.

[29] BHAYANA R，KRISHNA S，BLEAKNEY R R. Performance of ChatGPT on a radiology board-style examination：Insights into current strengths and limitations[J]. Radiology，2023，307(5)：230582.

[30] LU Y，LIU Q，DAI D，et al. Unified structure generation for universal information extraction[J]. arXiv：2203.12277，2022.

[31] ZHAO L，ZENG W，XU W，et al. Give the truth：Incorporate semantic slot into abstractive dialogue summarization[C]// Findings of the Association for Computational Linguistics：EMNLP 2021. 2021：2435-2446.

[32] LIU Z Y，SHI K，NANCY C. Coreference－Aware Dialogue Summarization[C]// Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2021.


**LIN Haonan**，born in 1998，M. S.，is a member of CCF (No. A1832G). His main research interests include natural language processing，smart healthcare，etc.

**TAN Hongye**，born in 1971，Ph. D，professor，is a member of CCF(No. E200022704M). Her main research interests include natural language processing，smart healthcare，smart education etc.