**CCKS 2020 技术评测任务书**

# 蕴含实体的中文医疗对话生成

# 任务描述

目前，我国每千人口执业医师数仅为2.44人，医疗资源分配不均、医生工作辛苦、患者就诊等待时间长等问题长期存在。面向智能问诊的医疗对话系统能够实时与患者进行沟通，收集患者信息，从而提升临床诊断的效率、减轻医生的负担、提高复诊积极性，有着广阔的应用前景[1,2,3]。本评测任务为蕴含实体的中文医疗对话生成，主要涉及对话系统构建、自然语言生成、领域知识融合等方向。

任务具体定义如下：给定医生和患者交流的前K句对话历史H={$X\_{1}, X\_{2},X\_{3},…,X\_{K}$}，其中$X\_{K}$为患者当前的对话语句，并假定医生的下一句回复$X\_{K+1}$包含标注的实体列表E={$e\_{1}, e\_{2},e\_{3},…,e\_{K}$}，模型要求根据对话历史H预测出下一句回复$X\_{K+1}$。同时，生成的回复中应当包含尽可能准确的医疗实体（E中的实体）。本任务基于一个带有实体标注的医疗对话数据集MedDG[1]，涉及12种消化内科相关疾病，包含超过1.7万个对话和38万个语句。MedDG中每个对话都标注了疾病、症状、程度、检查、药物五大类相关实体共160种。

任务的挑战如下：（1）医疗对话系统模型需要能够准确理解对话历史，生成与上文相关的对应回复。（2）训练集中，我们提供了五类实体标注，参赛者需要考虑如何利用这些信息来辅助模型训练。在验证和测试阶段，模型只能根据对话历史来推断回复中可能包含的实体。（3）任务提供的医疗对话数据有限，参赛者可以考虑利用开源的医疗领域相关语料进行模型预训练和迁移学习，例如数据集MedDialog [3]提供了3400万个无标注的中文对话。

[1] Liu et al. MedDG: A Large-scale Medical Consultation Dataset for Building Medical Dialogue System. ArXiv, 2020.

[2] Lin et al. Graph-Evolving Meta-Learning for Low-Resource Medical Dialogue Generation. In AAAI, 2021.

[3] Zeng et al. MedDialog: Large-scale Medical Dialogue Datasets. In EMNLP, 2020.

# 数据描述

本任务基于一个带有实体标注的医疗对话数据集MedDG，涉及12种消化内科相关疾病。其中训练集train.pk包含14863个完整对话，所有对话都标注了疾病、症状、属性、检查、药物五大类共160项相关实体。验证集和测试集中的样例均已删除实体标注，模型只能根据对话历史生成回复。每个验证和测试样例的正确回复中至少存在一个实体。实体类型具体介绍如下：

* 疾病：共12项，包括胃炎、肠炎、便秘等。
* 症状：共62项，包括腹泻、腹痛、腹胀等。
* 属性：共4项，包括时长、诱因、性质、位置。
* 检查：共20项，包括胃镜、肠镜、便常规等。
* 药物：共62项，包括奥美、吗丁啉、莫沙贝利等。

数据集中entity\_list.txt文件展示了完整的实体列表。

**训练集对话样例（带实体标注）：**

[{'id': 'Patient', 'Sentence': '胃部不适，第一天有痛感，后面就是胀，不拉肚子。就是胀气。请问是什么原因。（男，39岁）', 'Symptom': ['腹泻', '腹胀', '胃肠不适'], 'Medicine': [], 'Test': [], 'Attribute': [], 'Disease': []},

{'id': 'Doctor', 'Sentence': '您好，您的症状有多久了呢？', 'Symptom': [], 'Medicine': [], 'Test': [], 'Attribute': ['时长'], 'Disease': []},

{'id': 'Doctor', 'Sentence': '平时，有没有反酸嗳气，大便情况怎么样？', 'Symptom': ['打嗝', '反流'], 'Medicine': [], 'Test': [], 'Attribute': [], 'Disease': []},

{'id': 'Patient', 'Sentence': '您好，四天了。反酸到不明显。大便正常的。', 'Symptom': ['反流'], 'Medicine': [], 'Test': [], 'Attribute': [], 'Disease': []},

{'id': 'Doctor', 'Sentence': '四天前，有没有吃过凉的东西、辣的东西？有没有腹部受凉？', 'Symptom': [], 'Medicine': [], 'Test': [], 'Attribute': ['诱因'], 'Disease': []},

{'id': 'Patient', 'Sentence': '嗯，那天晚上12点吃了几块凉的饼。', 'Symptom': [], 'Medicine': [], 'Test': [], 'Attribute': [], 'Disease': []},

{'id': 'Patient', 'Sentence': '半夜就开始痛了。', 'Symptom': [], 'Medicine': [], 'Test': [], 'Attribute': [], 'Disease': []},

{'id': 'Doctor', 'Sentence': '您这个情况考虑是饮食所引起的胃肠功能紊乱，目前有消化不良的表现。用过什么药物没有。', 'Symptom': ['消化不良', '胃肠功能紊乱'], 'Medicine': [], 'Test': [], 'Attribute': [], 'Disease': []},

{'id': 'Patient', 'Sentence': '家里找了一些，胃苏冲剂，还有什么xx盐酸什么，但效果不明显。也没有持续吃。', 'Symptom': [], 'Medicine': ['胃苏'], 'Test': [], 'Attribute': [], 'Disease': []},

{'id': 'Doctor', 'Sentence': '建议这几天饮食上清淡一点，不要吃凉的东西，冷的东西。', 'Symptom': [], 'Medicine': [], 'Test': [], 'Attribute': [], 'Disease': []},

{'id': 'Doctor', 'Sentence': '可以口服莫沙比利+金双歧。', 'Symptom': [], 'Medicine': ['莫沙比利', '金双歧'], 'Test': [], 'Attribute': [], 'Disease': []}]

**字段说明：**

'id': 当前讲话者身份，'Patient'代表病人，'Doctor'代表医生。

'Sentence'：当前语句的具体内容

'Symptom'：当前语句中包含的症状实体

'Medicine'：当前语句中包含的药物实体

'Test'：当前语句中包含的检查实体

'Attribute'：当前语句中包含的属性实体

'Disease'：当前语句中包含的疾病实体

**验证、测试集样例（无标注）：**

**样例输入：**

[{'history': ['你好，肚脐周围隐隐作痛，不知道怎么回事（女，29岁）',

'你好，这种情况多长时间了？',

'两三天了。',

'隐隐作痛，疼一会就不疼了。',

'还有其他症状吗？恶心想吐吗。',

'没有。',

'是隐隐约约的疼吗。',

'食欲也好的，稍微有点腹胀。']}]

**样例输出：**

 ['可能是胃肠功能紊乱。']

# 评价指标

 本任务采用BLEU-1、BLEU-4和Entity-F1三项指标来评估医疗对话生成的质量。最终模型的得分为三个指标的算术平均值。

* BLEU-1和BLEU-4用于衡量模型生成语句与医生提供回复的文本相似程度，本任务评测主要考虑字符级别的一元组和四元组相似度，具体计算过程可参考论文[4]。
* Entity-F1用于衡量模型生成语句中涉及实体的准确程度，具体计算公式如下：

$$precision=\frac{生成语句中包含标注实体的个数}{生成语句中提及实体总数}$$

$$recall=\frac{生成语句中包含标注实体的个数}{医生回复中标注实体的总数}$$

Entity-F1$=\frac{2\*precision\*recall}{(precision+recall)}$

说明：在评测脚本中，主办方采用了字符串匹配方法来检测生成语句中出现的实体，字符串覆盖了训练集中出现的所有可能的实体表达。评测脚本不公开，参赛者如果需要自行计算Entity-F1，可以参考实体列表entity\_list.txt和训练集对话中提及的同义词。

[4] Chen et al. A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU

# 任务提交

本次评测将依托Biendata平台（https://www.biendata.xyz/）展开，请有意向的参赛队伍关注平台上的竞赛列表，通过平台注册报名参赛。

评测组织者会在Biendata平台上发布一个Baseline方案和实现，以及提交文件样例result.pk，供参赛队伍参考。选手将程序运行结果保存为result.pk文件进行提交，格式与组织者提供的result.pk相同。

本次任务将采取刷榜的方式，验证集发布后，允许参赛队伍多次向平台提交结果，参赛队伍可在测试集发布之前随时上传验证集的计算结果，系统会更新各队伍的最新排名情况；测试集发布后，允许参赛队伍多次提交测试集结果文件。具体提交次数限制将在Biendata上公布。

**最终提交文件要求**

每一个参赛队需提交的材料如下：

1. 测试集结果文件result.pk
2. 相关代码及运行描述文档（非评测论文，评测论文撰写要求见CCKS 2020官网）

以上三个文件需在任务提交截止日期前发送至邮箱ccks2021\_mdg@163.com。邮件的标题为：“CCKS-MDG-参赛队名称”，例如“CCKS-MDG-XXX队”。

代码及其文档需打包成一个压缩文件（tar，zip，gz，rar等均可），用code.xxx命名，要求提交所有的程序代码及相关的配置说明，**程序应当可以运行且所得结果与result.** **pk相符**。本任务允许使用开源语料和预训练模型，但禁止使用未开源数据。如果参赛者使用了额外资源，要求做出说明并提供资源文件或地址。

# 时间安排

时间安排初定如下，后续如果有调整，会在讨论组中更新说明。

• 评测任务发布：3月20日

• 报名时间：3月20日—7月15日

• 训练及验证数据发布：4月1日

• 测试数据发布：7月15日

• 提交测试结果：7月20日

• 评测论文提交：8月5日

• CCKS会议日期(评测报告及颁奖)：8月18—21日

# 参赛奖励

本评测总奖金30,000元，由腾讯天衍实验室全额赞助，并由CCKS专委会为任务前三名颁发证书。奖金设置如下：

第一名：15,000元

第二名：10,000元

第三名：5,000元

# 组织者

梁小丹 中山大学 xdliang328@gmail.com

唐鉴恒 中山大学 sqrt3tjh@gmail.com

刘文阁 中山大学 kzllwg@gmail.com

郑冶枫 腾讯天衍实验室 yefengzheng@tencent.com

赵瑞辉 腾讯天衍实验室 zacharyzhao@tencent.com

刘亚飞 腾讯天衍实验室 davenliu@tencent.com

陈曦 腾讯天衍实验室 jasonxchen@tencent.com

组织单位：中山大学、腾讯天衍实验室

# 任务交流平台

任务交流平台采用qq群的方式，所有报名参赛的参赛队要求有一名成员加入到该讨论群中。后续所有有关本任务评测及数据的相关说明和通知将只在讨论群中发布和交流，不再另行发邮件或更新官网通知。请所有参赛队务必在报名成功之后加入该讨论群，qq群号为603206031。

