

标注规范

第一部分：标注流程图解

Step1: 进入 121.41.9.10 并登录

医疗对话数据标注

登录

用户名

User Name

密码

Password

登录

Step2: 进入首页，可以看到已完成标注文本个数和待标注文本个数；请仔细阅读保密协议，点击“同意保密条款”进入下一页。

复旦大学 大数据学院

查看已标注文本

查看诊疗报告

查看标注说明

退出

已完成标注文本个数

0

待标注文本个数

5

保密协议

鉴于应对此等秘密信息承担的相关保密责任，本人在此同意严格遵守以下各条款：

1. 本人同意标注平台向本人披露的全部信息，包括但不限于病例自述、症状诊断、医患对话文本、标注结果、计算机程序、源代码等，不限传达方式及是否为草稿或最终确认稿，无论上述信息是否与过去、目前、和未来的商业活动、研究、生产设计或发展、人员或商业合作活动相关，均为归复旦大学大数据学院所有的秘密信息，以及复旦大学大数据学院从第三方处获得但应承担保密责任的信息，统称为“秘密信息”。
2. 对于上述秘密信息，本人同意：1) 应有意地保护此等秘密信息，并采取所有必要的保密措施；2) 不向公众披露此等秘密信息或向第三方披露此等秘密信息；3) 未经复旦大学大数据学院书面同意不得在任何情形下使用此等秘密信息；4) 不得对秘密信息进行复制、逆向工程、反编译、试图破译源代码及潜在信息。
3. 本人如应法院或其他政府部门的命令或要求而进行披露秘密信息，本人将会事先书面告知复旦大学大数据学院并回应所有可行的保密要求。
4. 本人将采取合理的防范措施保护此等秘密信息。未经复旦大学大数据学院书面同意，我方将不会以任何形式对秘密信息进行复制（含以纸张或软件或电子文档形式）。
5. 如此等秘密信息未经授权而被使用或披露，或本人有任何违反本协议的行为，本人将立即书面告知复旦大学大数据学院，并会以各种形式协助其收回秘密信息。

同意保密条款

Step3: 进入标注说明界面，点击“下载标注规范”可下载本文档，点击“开始标注”进入标注主页面。



Step4: 进入标注主页面，开始标注。



标注演示:

1) 选择每句话的话语行为:

Step5: 点击左侧的“查看已标注文本”，可对已标注文本进行修改，**请仔细核对是否有“未标记”动作以及标记实体是否准确。**



#	已完成标注文本	动作	标记实体	操作
10626355_1	你好	其他		修改
10626355_2	在吗	其他		修改
10626355_3	这个粑粑正常吗?	提问-症状		修改
10626355_4	从你发的图片看，这个孩子的大便有点稀，大便当中有粘液	告知-症状	症状：稀，粘液	修改
10626355_5	不正常的话和吃的奶粉有关吗	提问-病因		修改

Step6: 点击左侧“查看诊疗报告”，可对已书写的诊疗报告进行修改。



第二部分：标注说明

您将看到一段医生和患者的对话文本，包含疾病、患者提问和对话内容。您需要标注出以下三个内容：每句话所对应的话语行为、句子内部记录（包括实体识别和实体规范化）、每个病例的诊疗报告。具体说明如下：

（一）话语行为标注

每句话都有其对应的话语行为（dialogue act），话语行为一共包含十六类，分别是提问-症状、告知-症状、提问-病因、告知-病因、提问-基本信息、告知-基本信息、提问-已有检查和治疗、告知-已有检查和治疗、提问-用药建议、告知-用药建议、提问-就医建议、告知-就医建议、提问-注意事项、告知-注意事项、诊断、其他。

- （1）提问-症状：与症状有关的提问；
- （2）告知-症状：与症状有关的回答，也包括患者的肯定或否定回答；
- （3）提问-病因：询问症状和疾病产生的原因；
- （4）告知-病因：解释症状和疾病产生的原因，通常有“导致”、“引起”、“造成”等词；
- （5）提问-基本信息：基本信息包括患者个人信息、生活习惯（饮食习惯等）、既往病史、家族情况等；
- （6）告知-基本信息
- （7）提问-已有检查和治疗：已有检查和治疗包括做过的检查项目、化验单情况和药物治疗等，该动作必须由医生发出；
- （8）告知-已有检查和治疗：该动作必须由患者发出；
- （9）提问-用药建议：与药物及用药方式相关的提问；
- （10）告知-用药建议：与药物及用药方式相关的回答；
- （11）提问-就医建议：与检查、就医建议相关的提问；
- （12）告知-就医建议：与检查、就医建议相关的回答；
- （13）提问-注意事项
- （14）告知-注意事项
- （15）诊断：①重要信息，通常包含“考虑是”、“应该考虑”、“认为是”等词语，例如“你这个情况需要考虑是消化不良”；②医生通过已有检查报告等信息给出的判断也归为此类，例如“白细胞偏高”。
- （16）其他：①问候、感谢、与病情无关的句子；②空句子，表明该句为图片；③

部分条件语句表示后续的判断，例如“如果查出来有炎症，就需要进行抗感染治疗”；④医生的一些解释，例如“咳嗽是人体清除呼吸道内的分泌物或异物的保护性呼吸反射动作”。

请注意：

- 如果出现两者以上的话语行为，则根据实际情况判断主要的话语行为，无法判断重要性时，则根据就近原则选择第一个分句的话语行为。
- 如果患者提问“该怎么办”、“怎么治疗”，无法明确具体的提问目的，根据上下文选择。
- 如果不确定动作类别，可以结合上下文进行判断，如果发现仍然不属于上述类别，或者对诊疗结果没有影响，可归为其他。
- 有些能反映症状的基本信息需要归为症状类，例如疾病为“小儿消化不良”时，“每天大便几次？”的话语行为是“提问-症状”。

（二）句子内部记录标注

句子内部记录标注分为两大部分：**实体识别以及实体规范化**。

1、实体识别

实体识别是指识别出句子中重要的词组，如症状、药品名、药物类别、服用方式、检查、操作、注意事项等，并对每句话的每个字打上相应的 BIO 标签（B：实体开始，I：实体中间，O：其他文字），需要识别的实体一共包含七类：

（1）**症状**：症状实体之前已经人工标注了，因此本次标注时，与症状有关的，已经标注的内容，不改；如果有新的与症状有关的内容，也不增加。

（2）**药品名**：指具体的药品名称，例如“我建议你服用妈咪爱”，其中“妈咪爱”就是药品名；

（3）**药物类别**：需与药品名区分，指一类药物，例如“抗菌药”、“抑制胃酸分泌药”等；

（4）**服用方式等**：包括用药剂量、用药频次、口服、饭前饭后等相关文字；

（5）**检查**：与检查建议相关的名词，既包括具体的检查项目，如血常规、尿常规、化验、胸片、支原体检查等，也包括宽泛的建议，如就医、检查、去医院等；

（6）**操作**：与医疗操作相关的名词，例如输液、雾化、挂水等；

（7）**注意事项**：如喝热水、少食多餐等。

2、实体规范化

实体规范化是针对“**检查**”以及“**注意事项**”相关的实体提出的。具体说明如下：

(1) 规范化“检查”相关的实体时，如果规范化的词与 BIO 标注出来的词不一样的话，就在这一实体开头（6:B-检查）的下一行写上规范化名词（只需写在一个表格中即可）；否则，则不需要填上规范化词。例如原文中“小便常规”应该规范化为“尿常规”，原文中“尿常规”就不需要填写。

(2) 规范化“注意事项”相关的实体时，如果该词是否定的，则在这一实体的开头（8:B-注意事项）下一行填写“X”，若该词是肯定的，则不需要填写。例如“避免食用海鲜辛辣刺激的食物”，标注为“O O O O B-注意事项 | B-注意事项 | B-注意事项 | O O O”，其中，每个 B-注意事项下面都规范化为 X；“别着凉肚子”，标注为“O B-注意事项 | O O”，规范化为 X；“要多喝热水”，标注为“O O B-注意事项 | |”，不需要规范化。

请注意：

- 初始化的 BIO 标签已包含了部分信息，包括人工标注的**症状实体（这部分不需要修改）**以及**机器自动匹配的药品名和检查名（这部分不准确，需要进一步核实修改，尤其是将药物类别从药品名中分离开）**。
- BIO 标注时要尽量标注出关键词，不标无关的词。
- BIO 标签对应如下：0：O，1：I，2：B-症状，3：B-药品名，4：B-药物类别，5：B-服用方式等，6：B-检查，7：B-操作，8：B-注意事项。

（三）诊疗报告 report

每个病例需要输出一个诊疗报告，参考框架如下：

- (1) 主诉： 主要症状或体征
- (2) 现病史： 主要症状的描述（发病情况，发病时间）
- (3) 辅助检查： 病人已有的检查项目、检查结果、会诊记录等
- (4) 既往史： 既往的健康状况、过去曾经患过的疾病等
- (5) 诊断： 对疾病的诊断
- (6) 建议： 检查建议、药物治疗、注意事项