

之江实验室多模态情感数据集标注项目

一、项目总则

（一）解决方案要求

项目承接方需形成具体的、有针对性的解决方案纸质文件，包括公司背景、支撑条件、标注流程、质量监控、人员培训等内容，经委托方确认后执行。

（二）素材总量

素材库为 32 部电影、15 部电视剧（共计 544 集），存储空间 398G，在此素材库基础上切割中文对话片段。如果出现素材不足的情况，由项目承接方负责寻找新的素材，并经委托方确认后补充进入素材库。

（三）标注时长

最终标注的 utterance 总时长大于 1000 分钟。

（四）标签种类

1. 包含 27 种情感标签类型，并在标注数据集中较均匀分布。
2. 每种情感标签包含 1-5 级的强度标注。

二、标注规范

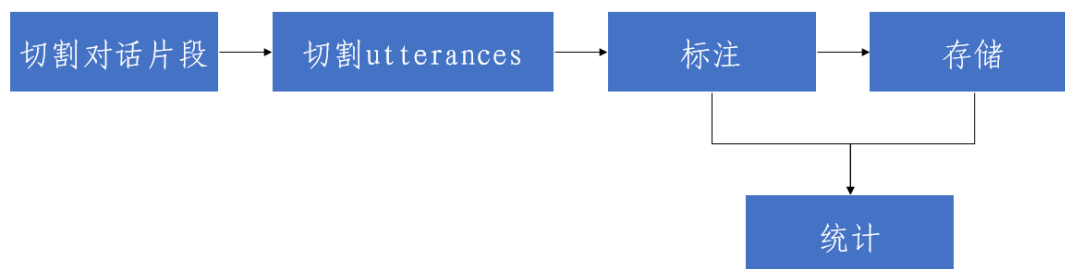


图 1：标注流程

（一）切割对话片段

1. 对话片段定义：在某一场景下，两个人对某一事件，连续交流产生的一段对话。
2. 根据人数、语音、明暗、背景声音等具体要求进行选取和切割。

（二）切割 utterances

1. utterance 定义：为表达某种情感，一个人连续说的若干句话。
2. 根据对话表达的连续性、完整性、情感性质、人脸表情等具体要求进行切割。



图 2：切割示例

(三) 标注

1. 由专业水平较高的 3 或 5 人独立开展标注工作, 通过多数投票的方式决定每一个 utterance 的最终情感标签; 将标注强度取平均值作为最终情感强度标签。
2. 标注内容: a) 每个 utterance 的文本 (人工将音频转换成文字并储存); b) 每个 utterance 视频片段和文本对应的情感标签; c) 人物的 id (0 或 1); d) 电影名称或电视剧名称+电视剧集数; e) 对话 id 和 utterance id; f) 开始和结束时间; g) 人物性别等; h) 情感强度。

对话 id	utterance id	人物 id	情感标签	电视剧名称或电影名称	电视剧集数	开始时间	结束时间	性别	文本	强度	包含谐音	包含历史典故、人名名言、谚语	情感类型
0	0	0	高兴	家有儿女	第100集	0:14:38	0:14:40	男	我考试及格了	3	0		0 基本情感
0	1	1	惊讶	家有儿女	第100集	0:14:40	0:14:47	男	你考试及格了?	5	0		0 基本情感

图 3: 标注示例

(四) 存储

1. 存储方法: 所有文件均需按照电影或电视剧的播放顺序依次存储, mp4 格式用于存储 utterance 片段 (如果不是 mp4 格式, 则转换成为 mp4); csv 格式用于存储 utterance 的标注内容。
2. 命名规则: dia(dia_id)_utt(utt_id).mp4/csv。

(五) 统计

标注数据统计信息: a) 不同对话人数量; b) 对话片段数量; c) utterance 数量; d) 男女老幼分布; e) 不同情感类型分布; f) 不同场景 (农村、城市、酒吧等) 分布; g) 每个 utterance

的情感标签一致性等。

三、验收标准

需同时达到以下标准为验收通过：

- （一）解决方案经双方确认同意并严格执行。
- （二）总任务量达到上述项目总则要求。
- （三）操作流程符合上述标注规范要求。
- （四）分别于标注时长达到 200 分钟（第一阶段抽检）、600 分钟（第二阶段抽检）、1000 分钟（第三阶段抽检）时邀请第三方介入抽检（需出具抽检合格证明），并邀请之江实验室专家团队介入评估，情感种类和强度合格率需达到 95%及以上，不合格的及时进行整改。
- （五）需于 2021 年 1 月 30 日前完成第一阶段抽检和评估，需于 2021 年 3 月 30 日前完成第二阶段抽检和评估，需于 2021 年 5 月 30 日前完成第三阶段抽检和评估，需于 2021 年 6 月 30 日前完成全部整改并交付。