NLP应用中数据治理遇到的困难及解决方案

彭嘉

小米集团技术委员会 NLP质量负责人

InfoQ 企业会员

企业数字化传播一站式服务

InfoQ 企业会员是为满足企业在中国开发者群体中的品牌曝光需求而推出的一款矩阵化资源包。可为企业提供包括“ 企业号服务 ”、“ 企业动态宣发 ”、“ 品牌展示通道 ”在内的多项专属权益与服务，助力企业高效触达开发者群体，提升数字化时代影响力。

企业号服务

深度触达 300 万中高端开发者

企业动态宣发

新媒体矩阵覆盖百万粉丝

品牌展示通道

线上平台 10 万+ 流量曝光

大纲

- 小爱智能助手介绍
- AI算法评估遇到的数据问题
- 线下线上评价结果不一致的解决方案
- 多次评测指标波动的解决方案
- 新探索与总结

1

小爱智能助手介绍

小爱语义标签结构

播放周杰伦的歌

小爱语义理解流程

离线算法指标

内容.音乐.按歌手查询

- 一级准确率/召回率
- 二级准确率/召回率
- 三级意图准确率/召回率

2

AI算法评估遇到的数据问题

NLP应用开发过程中常遇到的问题

离线算法指标好，线上表现就一定好么

如何持续地获得高质量的训练/评测集

上线前评测数据量级多大合适

算法指标有波动，到底应该相信哪一次

聚焦在AI模型上线前评测遇到的问题

离线算法指标好，线上表现不一定好

搜索推荐算法评测指标好，但是线上首条完听率并没有提升

多轮对话准确率从90%提高到93%，但是线上用户重说率没有显著降低

模型策略都没改动，数据集没变化，指标有波动

同一时间段多次评测模型，指标一直上下波动 .

10:05	95.17%
10:10	95.29%
11:03	94.98%

3

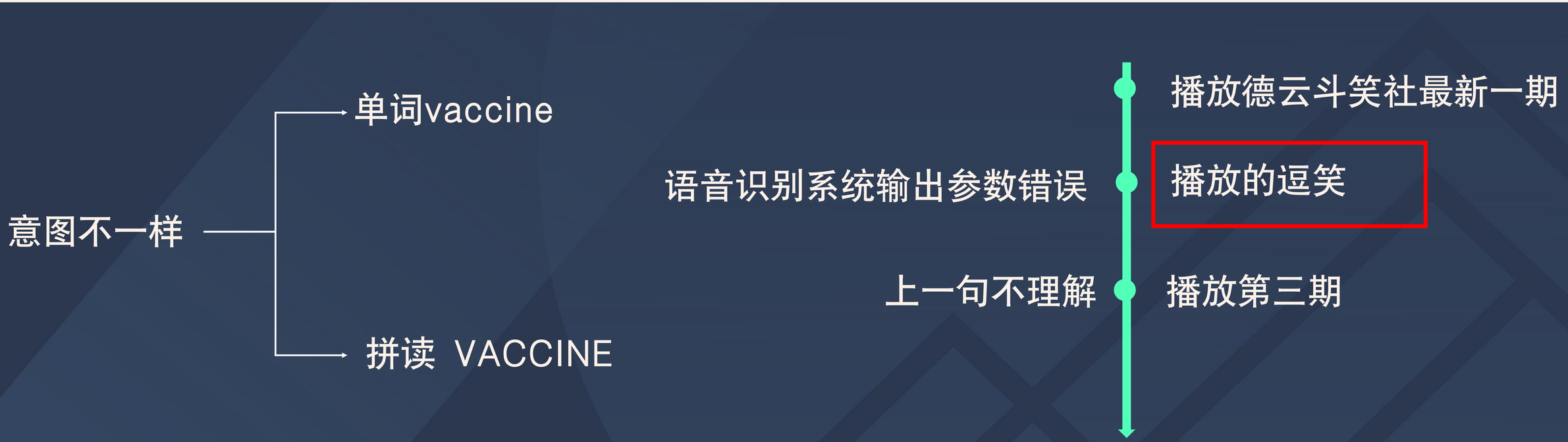
线下线上评价结果不一致的解决方案

表现不一致的因素拆解

- 评测环境难以评估多个系统相互施加的影响
- 线下和线上的评价体系不同
- 标注结果逐渐偏离普通用户认知
- 模型的训练和评测使用的是历史数据

评测环境难以评估多个系统相互施加的影响

唤醒 → 语音识别 → 语义理解 → 结果满足



线下和线上评价体系不同

算法团队

AUC　　F1值

Precidion　　Recall

业务团队

收听时长　　不满足重说率

...

用户感知

流畅程度

博学程度

换一种表达听不懂

...

滥用标签导致逐渐偏离用户认知

“安静”

- 静音（扬声器持续关闭）
- 暂停播放

模型的训练和评测使用的是历史数据

随着时间的推移，线上特征分布可能发生了变化，比如新冠，奥运会

线上线下表现不一致问题的解决方案

单一系统和全链路 自动化评测都要有

语音交互全链路的端到端评测，通过设置**环境路由**实现模块和整体的效果比较

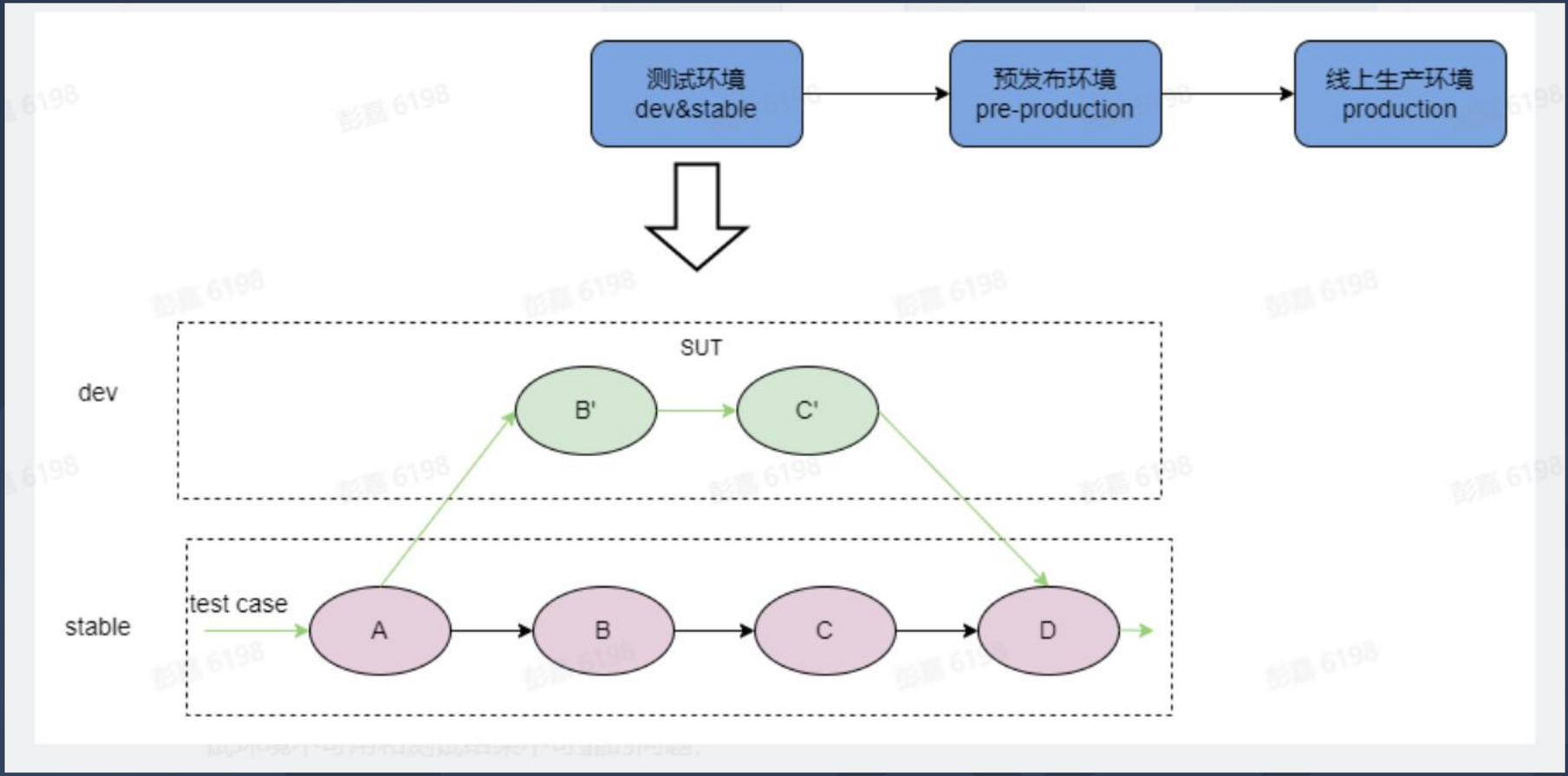
反思指标体系，警惕指标与用户问题脱节

深入了解用户，用 **新视角** 重新看待数据

承认随机性普遍存在

线下评测指标好，只能说明新模型或策略**大概率**比线上正在运行的好

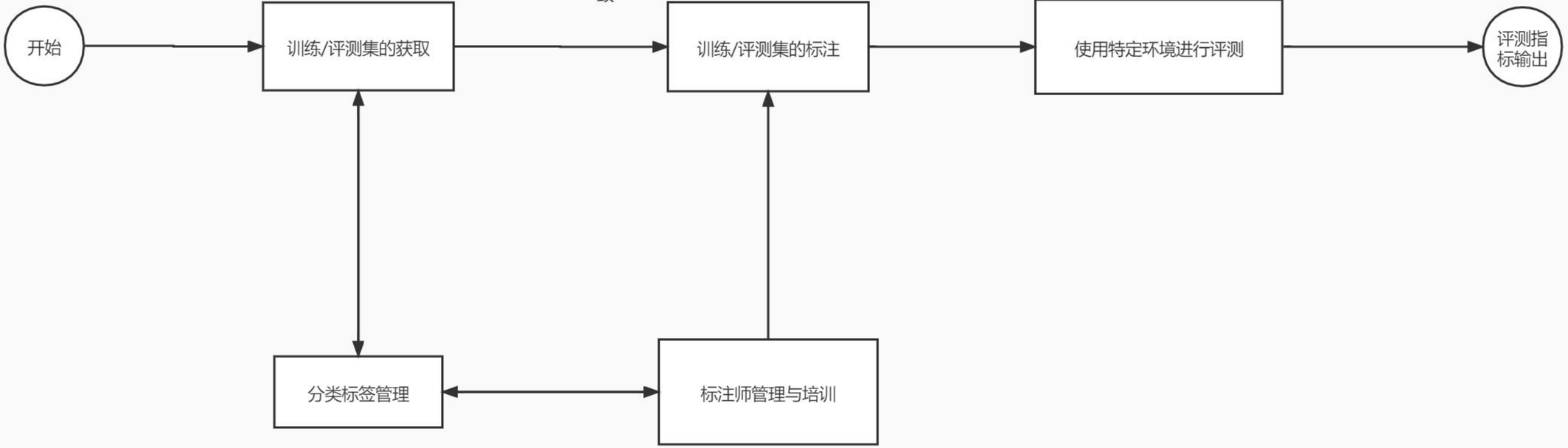
有环境路由功能的全链路评测架构



4

多次评测指标波动问题解决方案

影响评测指标的因素



难点：

· 冷启动阶段数据如何获取

· 正常运营后，采样数据是否符合现实的数据分布，是否足量

难点：

· 语义标注只考虑用户对话本身么

· 公司/组织对领域知识的特别规定和大众的通识不一致怎么办

· 人工标注的随机性导致相同用户对话标注结果不一致

难点：

· 服务是否稳定，吞吐量是否达标

· 评测环境与线上环境的数据隔离、与外部服务的鉴权

· 代码管理是否规范

难点：

· 怎么保证分类标签的正交

· 如何保证标注师更容易使用这些标签

· 如果无法避免少部分语义标签非正交，如何事先定义原则

难点：

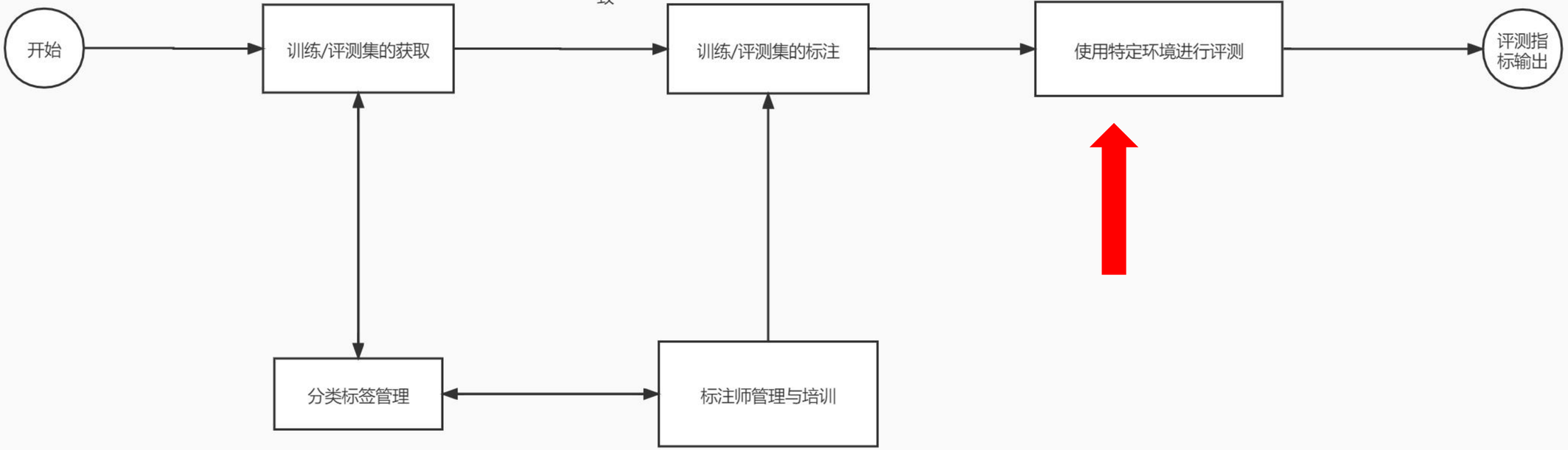
· 由于人员流动的原因，替补标注师的培养问题

· 如何评定标注师团队整体水平

· 如何检验单个标注师对分类标签认知和我们一致

· 对标注师效率和质量要求的平衡管理

指标的最终用户仅关注最后一个阶段：启动评测和指标输出



难点：

· 冷启动阶段数据如何获取

· 正常运营后，采样数据是否符合现实的数据分布，是否足量

难点：

· 语义标注只考虑用户对话本身么

· 公司/组织对领域知识的特别规定和大众的通识不一致怎么办

· 人工标注的随机性导致相同用户对话标注结果不一致

难点：

· 服务是否稳定，吞吐量是否达标

· 评测环境与线上环境的数据隔离、与外部服务的鉴权

· 代码管理是否规范

难点：

· 怎么保证分类标签的正交

· 如何保证标注师更容易使用这些标签

· 如果无法避免少部分语义标签非正交，如何事先定义原则

难点：

· 由于人员流动的原因，替补标注师的培养问题

· 如何评定标注师团队整体水平

· 如何检验单个标注师对分类标签认知和我们一致

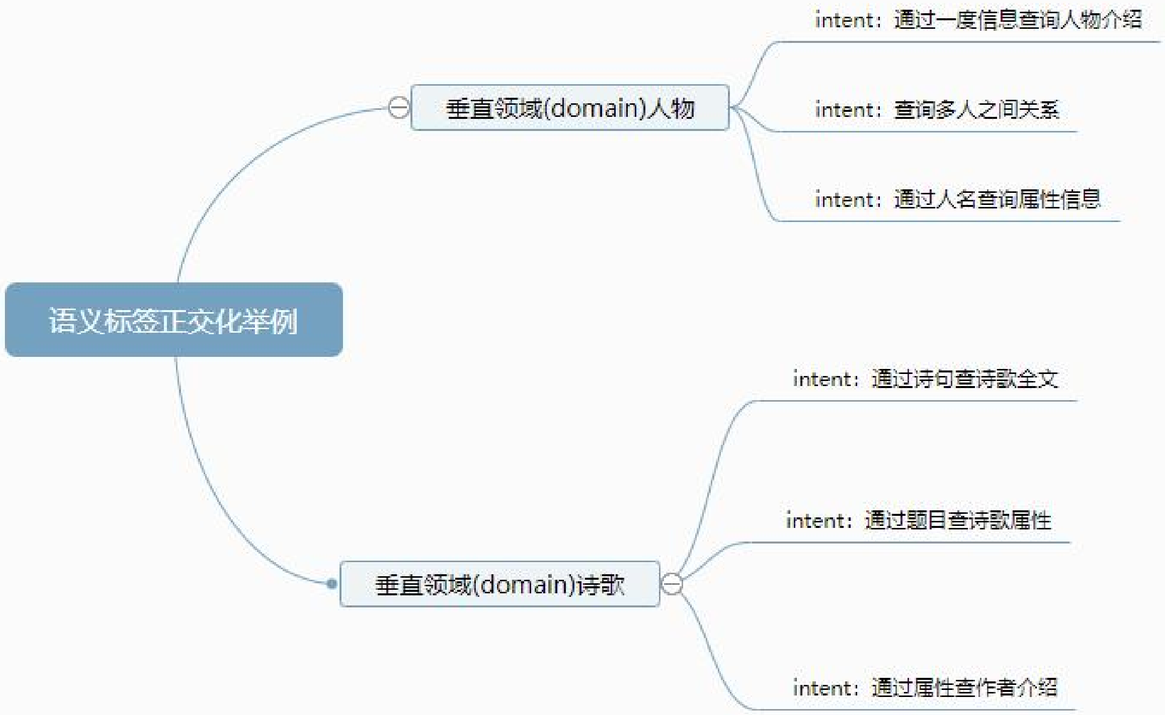
· 对标注师效率和质量要求的平衡管理

上文5 过程中涉及3 核心因素

解决指标波动-分类标签尽量正交

语义空间大，提前定原则

举例说明分类标签的正交



正交检查的原则：

将每个垂域领域的分类标签抽象归纳成**谓词结构**，将相似谓词结构短语用思维脑图连接一块。

案例：

诗歌垂域：通过属性查询作者介绍->通过属性查询人物介绍->通过一度信息查询人物介绍

VS

人物垂域：通过一度信息查询人物介绍

标签冲突时处理原则1

- 精细化运营的垂直领域语义标签优先于通用的垂直领域语义标签。

标签冲突时处理原则2

- 直接满足用户需求的语义标签优先于间接满足用户需求的语义标签。

小结-分类标签定义技巧

提前定义标注原则

正交避免标签冲突

冲突标签看谁优先

解决指标波动-指标误差率的计算

永远不可能把所有数据标注完，永远无法精确衡量模型对全量数据的准确率。

根据大数定律

随着样本容量n的增加，多次采样评测的准确率平均值接近于总体评测准确率。

所以需要多大的样本才能准确衡量模型

解决指标波动-指标误差率的计算

问题抽象为：总体的平均值和标准差未知的条件下，依据样本，对样本来自的总体参数进行估计

n:样本数

moe(margin of error)：多次评测指标误差率

σ :总体标准差

t：统计量在t分布下，你想要的置信度下的 t 值

$$\sqrt{n} \geq t \times \frac{\sigma}{moe}$$

解决指标波动-指标误差率的计算

通过除法，消除同类项

$$\frac{\sqrt{n1}}{\sqrt{n2}} = \frac{\frac{t*\sigma}{moe1}}{\frac{t*\sigma}{moe2}}$$

$$\sqrt{\frac{n1}{n2}} = \frac{moe2}{moe1}$$

指标的误差幅度受统计样本大小的影响，随着样本大小的增加，误差率也会降低

解决指标波动-指标误差率的计算

实验步骤：

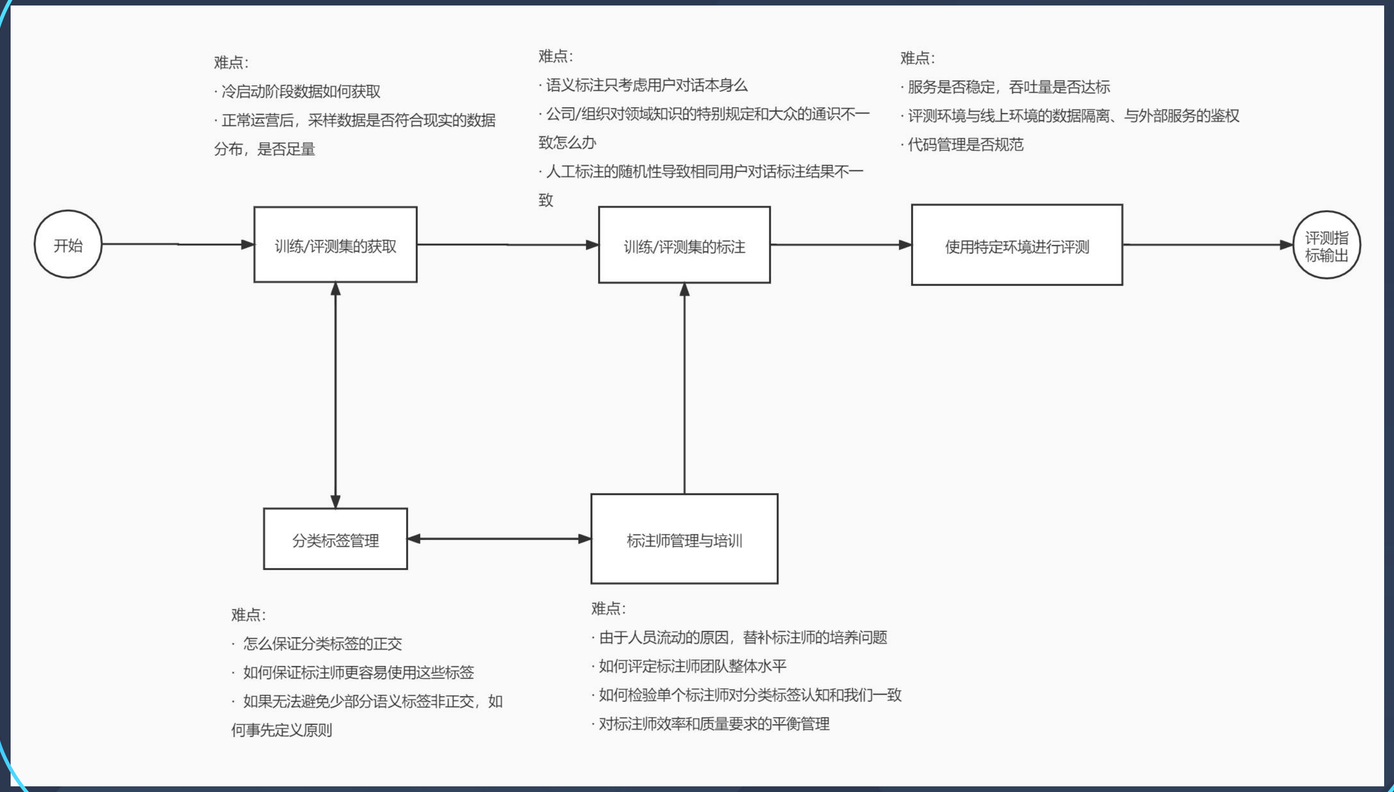
标注师的素质，分类标签的复杂度，评测环境，数据集的标注准确率在短期内不会有变化

1. 取一个固定样本量n1（不要太小），多次评测，计算评测的结果的平均值和标准差
2. 业务提出希望达到的理想误差率是多少，假设0.05%
3. 放入公式计算需要的样本量n2。

总结：

表示基于贵厂商现有**标注师**素质，分类**标签**的复杂度，评测**环境**，需要n2样本，才能达到业务需求。

小结-减小指标波动



指标是复杂流程的最终产物

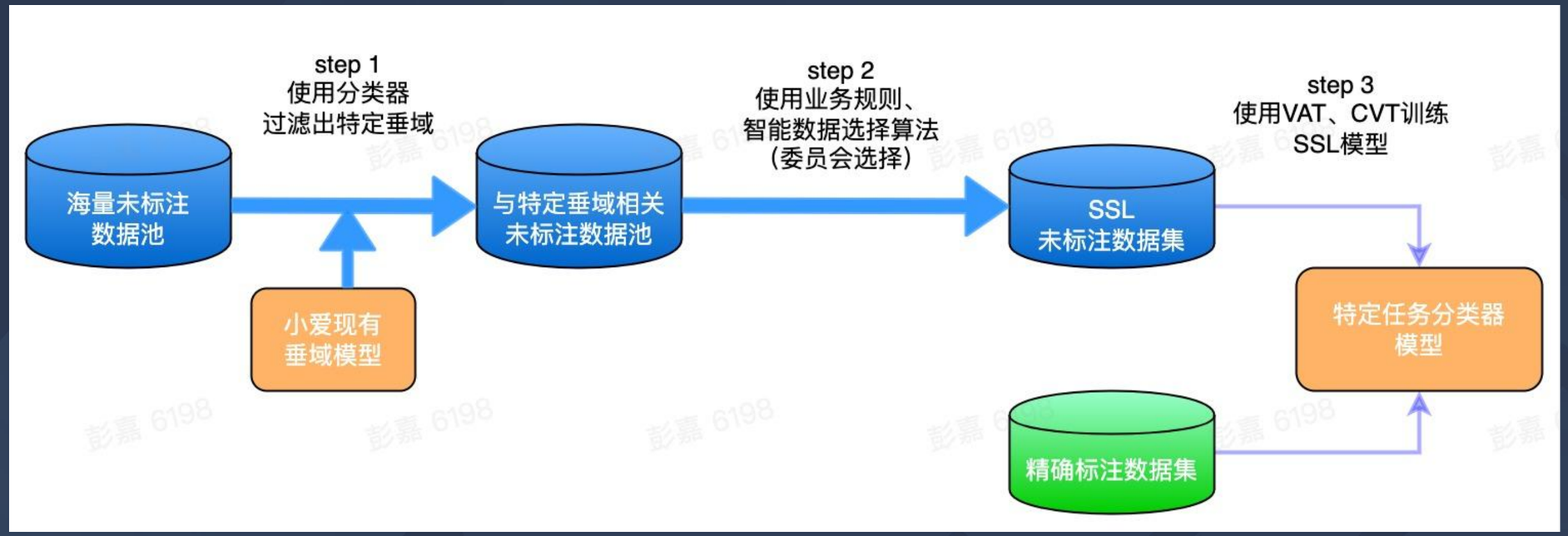
不盲目相信经验数据

用实验确定本公司的评测量级和指标误差

5

新探索与总结

探索使用半监督学习解决新业务标注数据少



总结

要想算法研发的努力不付之东流

要想模型策略的迭代能够真正推动业务

保证分类标签符合质量需求

保证指标体系符合用户感知

保证数据符合业务实际状况

THANK YOU.

永远相信美好的事情即将发生

彭嘉

小米・NLP质量负责人

小米・人工智能部意图定义委员会主席