


DataFunSummit

语音交互中的无效query识别

崔世起
小米人工智能部-小爱

目录

CONTENTS

PART 01. 无效query介绍

用户query的类型划分

	Query类型	Query示例	意图
有效query (可以结果满足)	单轮意图明确	*打开客厅空调*	智能家居
	场景意图明确	*千与千寻 (前台应用:music)*	音乐
		换一个 (上一轮：讲个笑话)	笑话
	多意图	*灰姑娘*	音乐|电台|视频
无效query (无法结果满足)	非人机交互	*你没打完呢*	无
	意图不明	*让谁先待会儿下雨*	未知

非人机交互query

定义　query不是用户给设备下发的指令

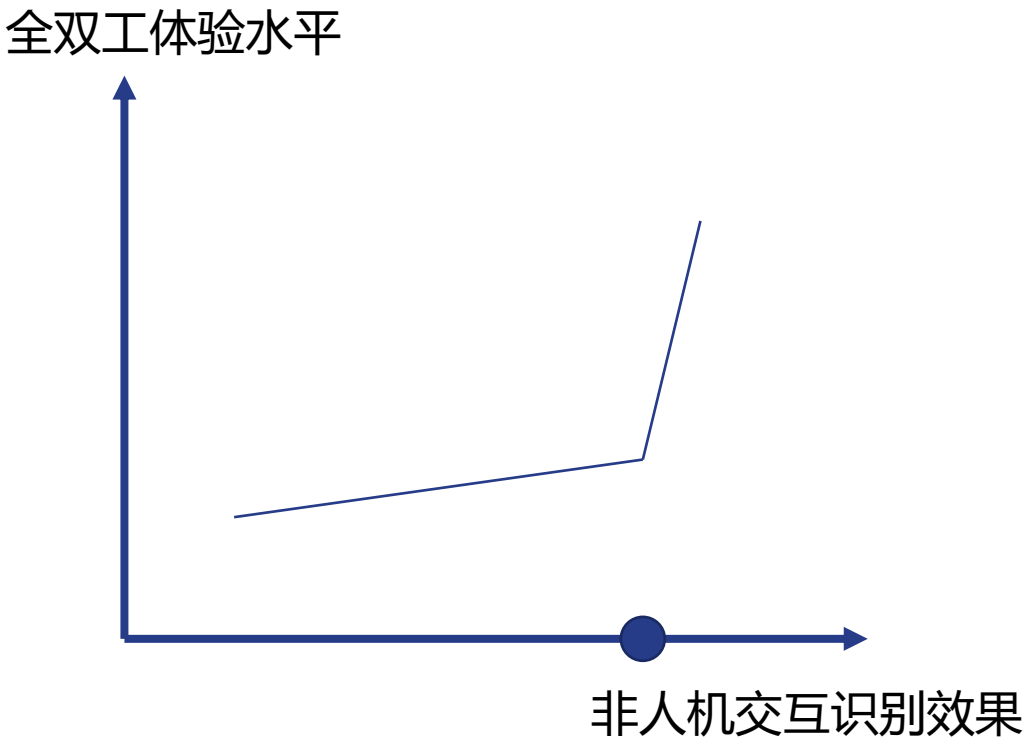
原因　误录入周围的人声

你没打完呢

那个跨年演讲不错

影响　乱响应，打扰用户

影响全双工连续对话体验



意图不明query

定义　query是人机交互指令，但根据query无法判断用户意图

类型　乱序无意义　*播放吴楚丽江穿个*

表达不完整　*帮我把手机*

Query意图比较模糊　*召唤、天空*

影响　误落入精品垂域，答非所问

非人机交互 + 意图不明　占比：5%-20%

无效query体验优化

非人机交互：识别 + 不响应

意图不明：识别 + 兜底回复或者引导澄清

PART 02. 非人机交互识别

非人机交互识别-难点

信息不完备

- 判断是否人机交互，需要多维度的信息（声音、视觉等）
- 只靠音频信息会有很大的歧义性

语音变化的多样性

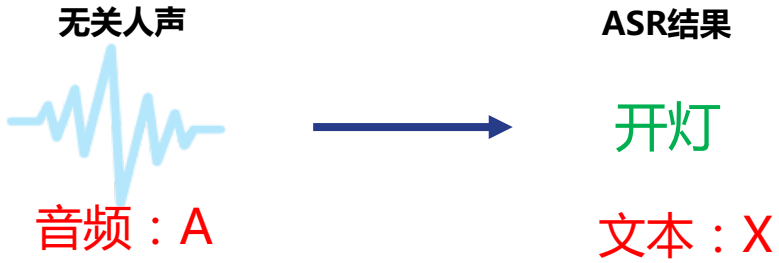
- 同一句话，由于语气、语调、语速、音色的不同，产生不同的音频
- 数据样本难以覆盖每种类型的语音

鸡尾酒会效应

- 嘈杂环境下的有效指令识别

非人机交互识别-问题建模

1. Language understanding Or Speech understanding



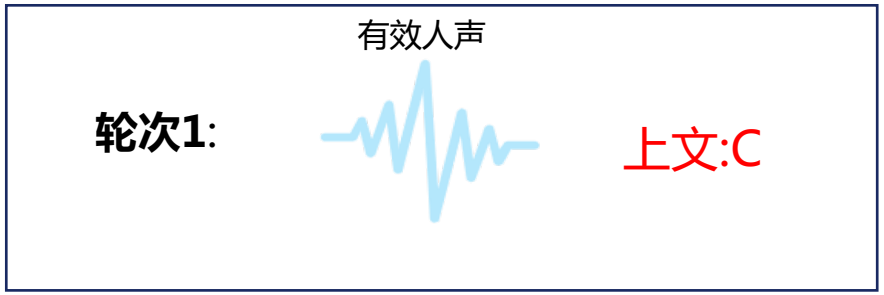
$\mathcal{P}(Y|X)$ ×

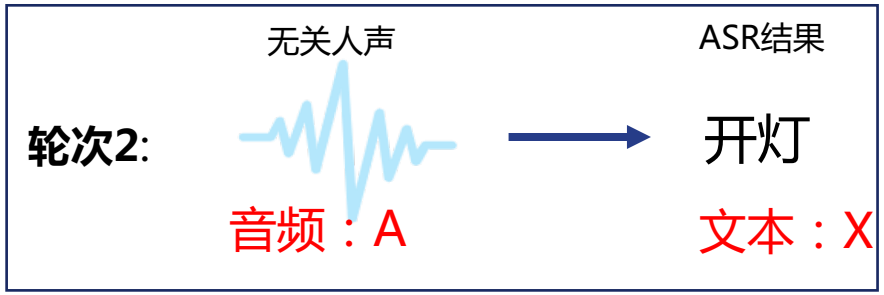
- 缺失了音频中包含的识别信号
- ASR错误传播

$\mathcal{P}(Y|A)$ ✓

非人机交互识别-问题建模

2. 是否需要依赖上文





$\mathcal{P}(Y|X)$ ✗

- 缺少音频特征
- ASR错误传播

$\mathcal{P}(Y|A)$ ✓

$\mathcal{P}(Y|X,C)$?

- 轮次之间弱依赖
- 增加上文依赖，建模难度增大

非人机交互识别-解决方案

建模

针对单轮语音的二分类

关键任务

数据集构建 + 特征和模型设计

非人机交互识别-数据集构造

非人机数据标注成本很高

听音频标注，标注10万条样本耗费100人日。

提升数据标注质量

1. 存在大量的模糊样本，需提高label一致性
   - 详细的标注规范，针对各类型音频提供示例
   - 多人标注验证

数据质量的提升 -> 模型效果的提升

非人机交互识别-数据集构造

样本挖掘

1. 提升样本的多样性
   - 随机采样
   - 正样本挖掘：基于ASR置信度、基于误唤醒检测
2. 提升样本的有效性
   - 挖掘困难样本：模型打分置信度低
   - 挖掘错误分类样本

非人机交互识别-模型分析

语音特征

频谱 优于 mfcc、fbank特征

加入通过声学信号处理获取的特征没有提升

语音Encoder

CNN -> CNN+LSTM+ATTENTION

CNN是个很强的baseline

非人机交互识别-模型分析

文本Encoder

CNN、TRANSFORMER、BERT效果差异不明显

语音Encoder和文本Encoder的融合

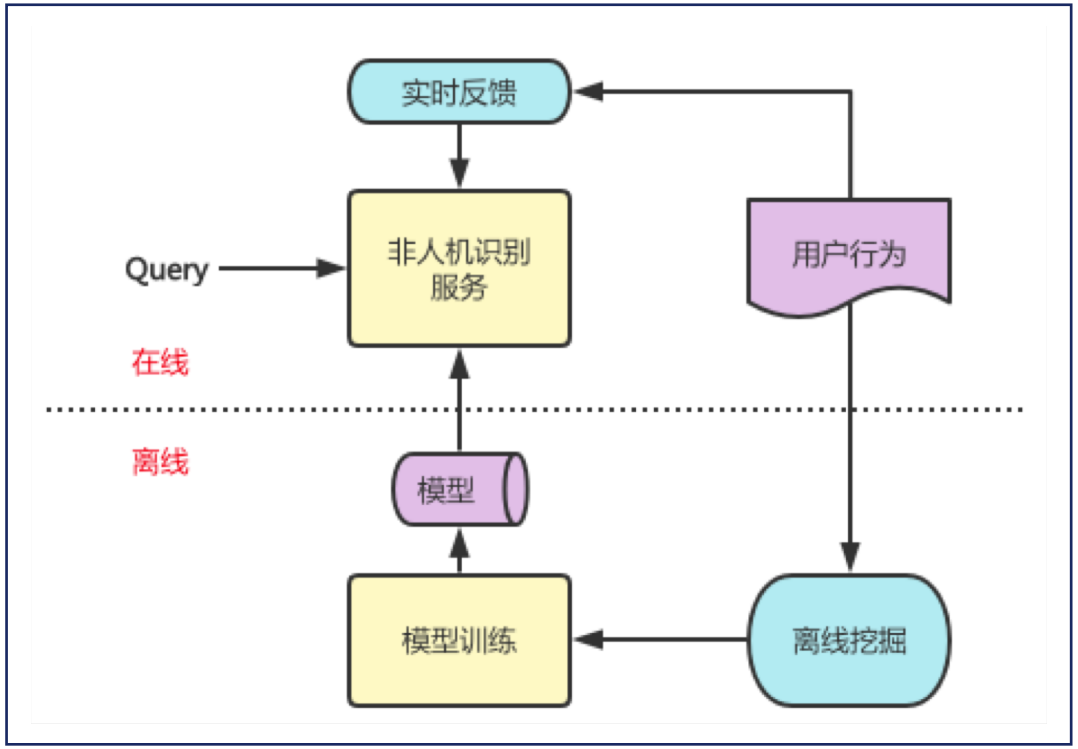
concat 优于 attention

用户行为反馈

用户反馈类型

误拒识反馈：拒识后重复说

欠拒识反馈：用户说“闭嘴”



反馈生效方式

在线：动态调整策略

离线：反馈数据进入模型迭代

个性化策略

引入Context：基于用户的历史行为和session

非人机交互识别-能力现状评估

疑问：基于语音的非人机交互识别，天花板在哪里?

评估方法：评估普通人在非人机交互识别上的平均水平。

评估对象：普通标注人员，未经过非人机交互标注的专业训练。

	标注人员1	标注人员2	标注人员3
准确率	84.03%	78.40%	93.21%
召回率	91.32%	95.85%	77.74%

结论：

普通人的识别准确率/召回率方差很大，平均F1值约为0.86。

目前在手机语音助手上已经接近普通人的水平。

PART 03. 意图不明识别

意图不明识别-问题类型划分

1. 乱序无意义

播放我在接桌子

关机后还会提醒说明天

你给我关闭台灯唯一听的为说

语法规范性

2. 表达不完整

眼泪是怎么

播放一首

明天上午8点

语义完整性

3. 意图模糊

安装包

成为

答案

意图强弱

解决方案：分而治之

乱序无意义识别

目标：将有序的query和乱序的query区分开

衡量文本有序性的指标：perplexity，通过语言模型计算

$$ppl(W) = P_{LM}(W)^{-1/N} = e^{CE\_loss(W)}$$

$-\log P(\text{厅}|\text{context})$

打开客厅空调

打开支付宝付款码

播放米小圈上学记

今天天气怎么样

有序

$-\log P(\text{揍}|\text{context})$

草莓灭绝揍人坝子

汤吃汤麦芽度杜鹃绚丽

放越来越难一别水

莱纳布朗灯

乱序

乱序无意义识别-如何得到更合理的perplexity

1. 足够多的训练数据，包含各种长尾知识、新词。

圣锤之毅的缺点	打工人
犹留正气参天地永剩丹心照古今	新冠肺炎
以雷霆击碎黑暗	网抑云

2. 足够大的语言模型：BERT、GPT

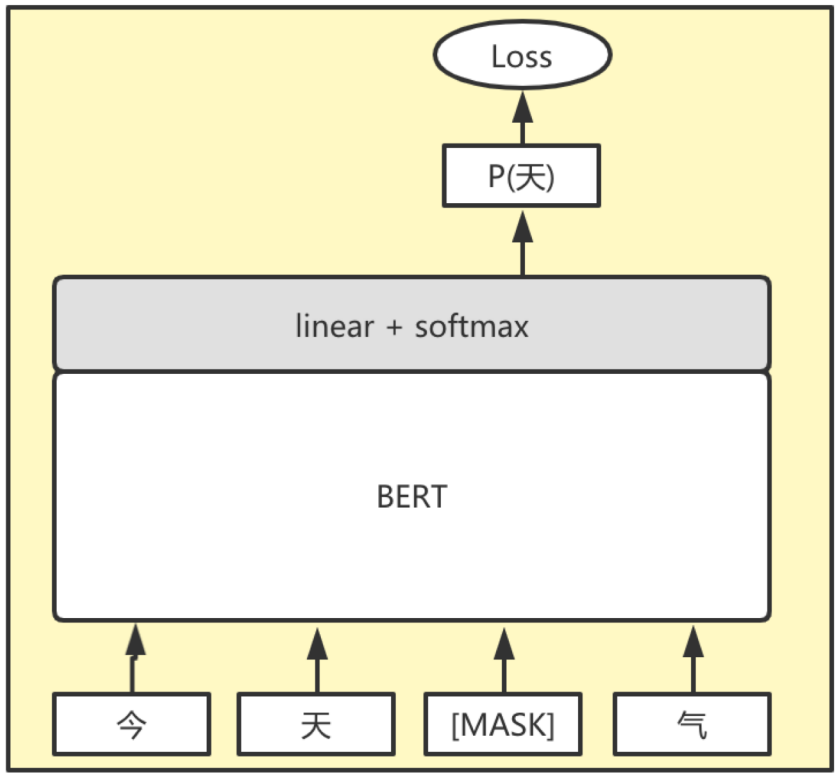
乱序无意义识别-语言模型方案

LSTM

$$P_{LM}(W) = \prod_{t=1}^{N} P_{LM}(w_t | w_1, \ldots, w_{t-1})$$

BERT

$$P_{MLM}(W) \approx \prod_{t=1}^{N} P_{MLM}(w_t | w_{\backslash t})$$

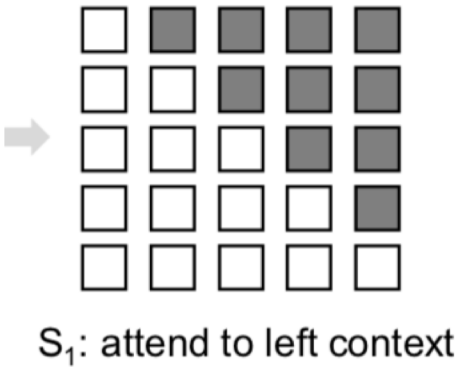


- Masked Language Model
- 训练：每个句子只Mask一个token
- 预测：依次Mask每个token，计算交叉熵损失
- 效果：同等Precision，Recall提升
- 缺点：预测阶段的时间复杂度太高

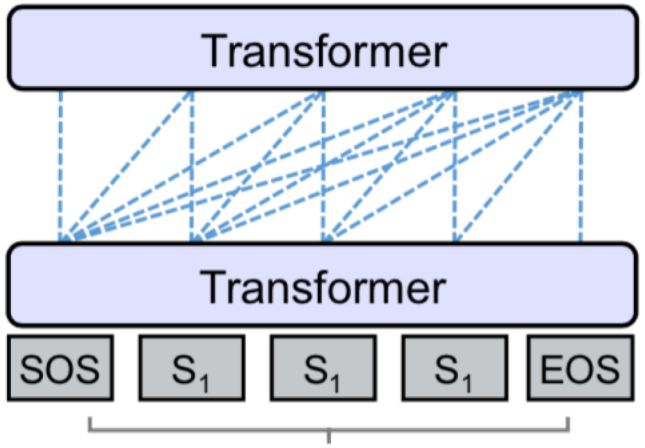
乱序无意义识别-语言模型方案

BERT

UNILM



S_1: attend to left context



- Left-To-Right LM
- Precision/Recall相当
- 预测复杂度降低N倍

乱序无意义识别-语言模型方案面临的挑战

有序和乱序边界的样本如何区分

四五六快进二十分钟	有明确意图
你给我关闭台灯唯一听	有明确意图

缺点：只用perplexity作为阈值，无法有效区分边界区域的混淆样本

解决思路：引入更丰富的特征，训练二分类，识别边界区域的正负样本

表达不完整识别

目标：判断一句话是否语义完整

播放一首

眼泪是怎么

今天是什么

周杰伦的

明天上午8点

建模

语言模型：$P(EOS|X)$ ，预测query下一个token是句尾的概率

二分类：$P(complete|X)$，预测query语义完整的概率

表达不完整识别

语言模型不好解决的case

倒装：放一首歌刘德华的（完整）

倒装：定个闹钟8点的（完整）

省略：一加一等于（完整）

特殊实体：播放我以为（完整）

前截断：成都的天气（不完整）

表达不完整识别

上下文相关的case

Q：明天早上八点（不完整）

Q：定个明天早上的闹钟 - A：明天早上几点?

Q：明天早上八点（完整）

建模方案

单轮分类：$P(complete|X)$，根据当前query判断是否完整

多轮分类：$P(complete|X, \text{context})$，判断当前query结合context是否完整

模型：BERT分类

小结

非人机交互识别

信息不完备的机器学习任务

基于语音和语义特征的神经网络模型

意图不明识别

乱序无意义识别和表达不完整识别两个任务

技术方案框架：语言模型 + 分类模型

THANKS !