QCon 全球软件开发大会
INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

BEIJING 2018

小Q机器人诞生之路

演讲者 / 王辉

主办方 Geekbang. 极客邦科技 InfoQ

自我介绍

- **王辉**，现就职于腾讯，社交平台部创新产品技术负责人，专家工程师；
- 近期主要关注人工智能和在线教育等创新产品的研发；
- 热爱技术，热爱跑步。

小Q机器人

AI技术、内容、硬件三大提升。

小Q机器人一代

2010

小Q机器人二代

2017

小Q机器人二代

https://item.jd.com/6422214.html

Agenda

聊天机器人

QQ空间聊天机器人


洋芋猫儿
明天有约会？万事俱备，试试输入"天气"吧
我们聊聊天吧
偶们很久都没聊过


全部动态
不一次遇上
只想说一句
我喜欢你好久了
无论别人有多好，我们都不要羡慕。
HarveyZhu，朱辉颖
今天19:59
美團好「敬業」
一次購買發四條 Notifications
送達之後，還提醒我去發圖


中国移动
下午2:36
沉迷
今天12:56
宝宝不开心了
迷爷特制的华为Mate 8 (4G)
浏览55次
对这条FEED,想说:
“有啥不开心的说出来让我开心开心”
发表
评论

技术成果

- 峰值每日聊天**2000w**次，用户**130w**（业界领先）
- 单用户会话轮次达到**15**次
- 应用多种业界前沿算法，并有多项创新
- 核心AI模型，算法实践结果做到**70%**的效果提升

技术方案

开域

海量数据 → 检索式 + 深度学习

核心场景：

- 聊天
- 推荐语

闭域

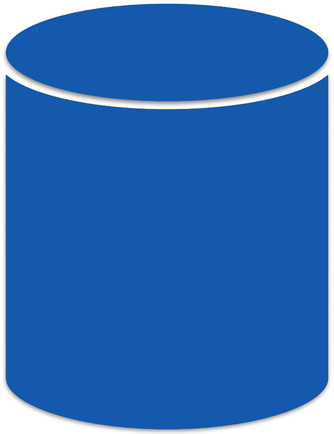
多样化数据 → 生成式

核心场景：

- 百科
- 知识问答
- 文字游戏

技术特色——海量数据

数据是人工智能的基石。

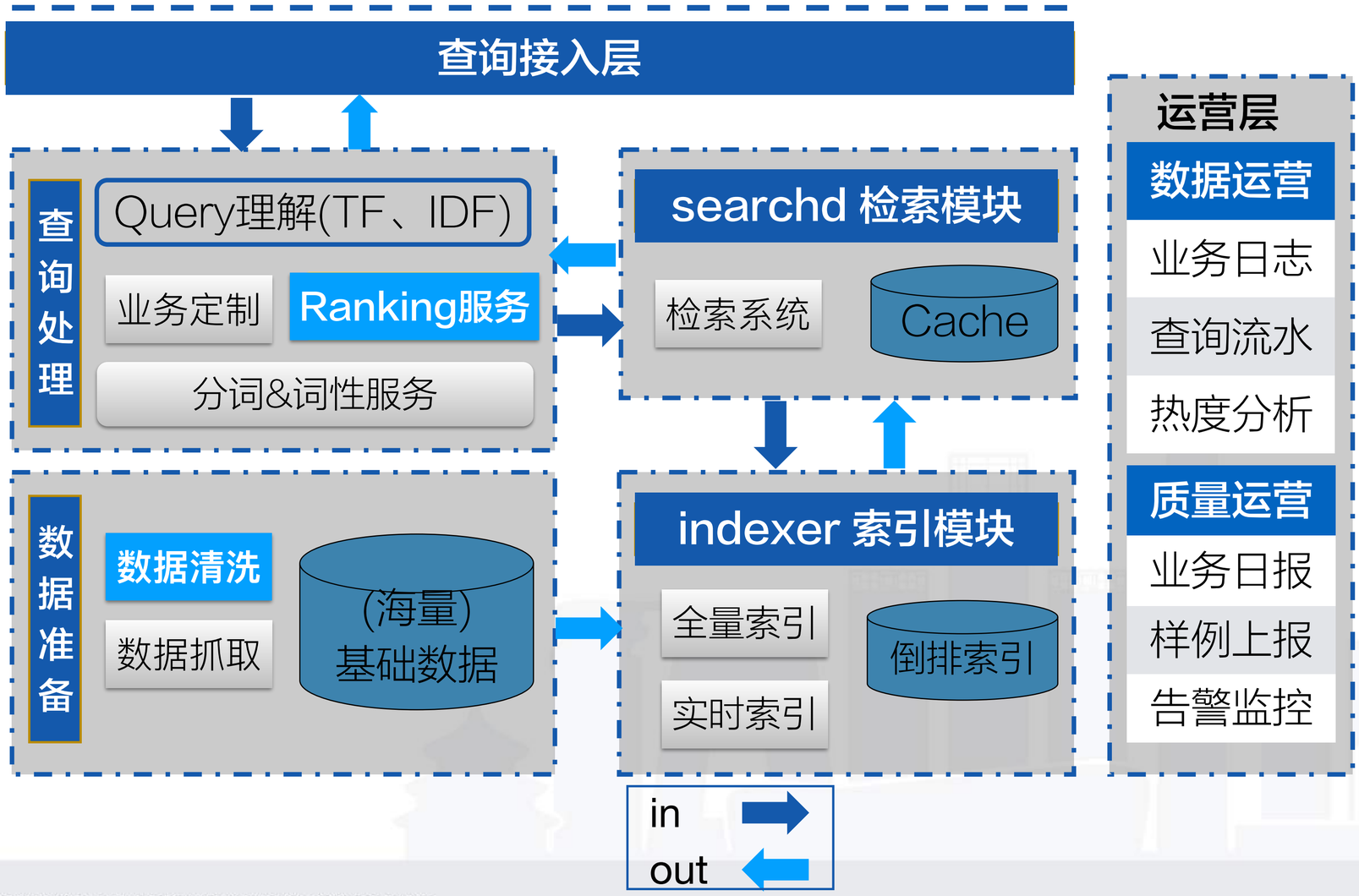


数十亿级公开问答对

技术特色——流动的海量数据

- 自我学习，机器人会“反思”（目前每天可以发现聊天内容约10%能继续学习到更好的回答）
- 开域转闭域，深度定制（多个闭域如八卦、新闻、影评等实时知识性数据每日更新，对话过程中的意图领域迁移跟踪）
- 新陈代谢，与时俱进（让聊天随着语料数据得到学习进步，整个数据清洗/检索索引/Rank模型适配流程化）

技术特色——高性能系统



Sphinx

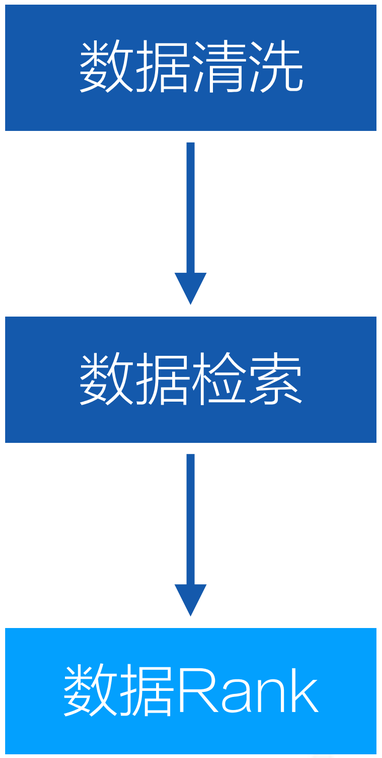
基于Sphinx改造的检索系统

约10亿的中文文档，规模超越知名的coreseek

分布式改造

单机性能优化算法稳定控制在单次计算耗时平均40ms.

技术特色——独家配方Rank模型



关于：最近加班频繁啊 的相关回答：

共检索到匹配答案：20

答	原问题
还在加？	最近加班比较频繁啊
有加班费吗？	最近加班比较频繁啊
还在加？	最近加班比较频繁啊
有加班费吗？	最近加班比较频繁啊
是滴，有点小忙	最近星期天加班比较频繁嘛。
嗯最近加班是正常的，不加班不正常	加班挺频繁的啊
，不是你的错，飞不起来可就是你的不对了	晚上加班太频繁了
还在加？	最近加班比较频繁啊
有加班费吗？	最近加班比较频繁啊

原创baseline基准算法

理论基础是条件概率，通过统计海量数据，当人问出一个什么词的条件下，人类回答什么词汇的概率比较高，得出聊天对话中语义上共现的规律。

Q: 今天踢球踢得很累

A1: 休息下

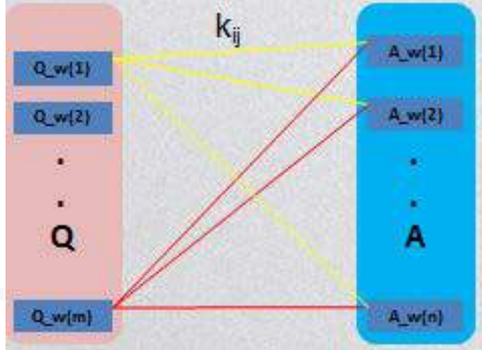
A2: 你还会足球啊?

A3: 下次叫我一起，球场见

$$P(A|Q)=P(A\_w|Q\_w)\ldots\ldots(1)$$

$$P(A\_w|Q\_w)=\sum_{i=1}^{m}\sum_{j=1}^{n}k_{ij}*P(A\_w(j)|Q\_w(i))\ldots\ldots(2)$$

$$P(A\_w(j)|Q\_w(i))=\frac{P(A\_w(j),Q\_w(i))}{P(Q\_w(i))}=\frac{Count(A\_w(j),Q\_w(i))}{Count(Q\_w(i))}\ldots\ldots(3)$$



问题中含有m个词，每个词表示成Q_w(i)
答案中含有n个词，每个词表示成A_w(j)

可当成二部图来看，词对之间权重相同，取值1/(m*n)

算法	P准确率	R召回率	F-measture	效果提升
词共现算法	0.532	0.265	**0.354**	-----

POS-IDF词向量权重模型

很多词语对确实在QA里共同出现了，但是他们并不是语义上有对话关系，所以需要修正这个条件概率的公式。

Q: 今天**踢球**踢得很累

A1: 休息下

A2: 你还会足球啊?

A3: 下次叫我一起，球场见

原理:

IDF文档倒频率体现词语关键性

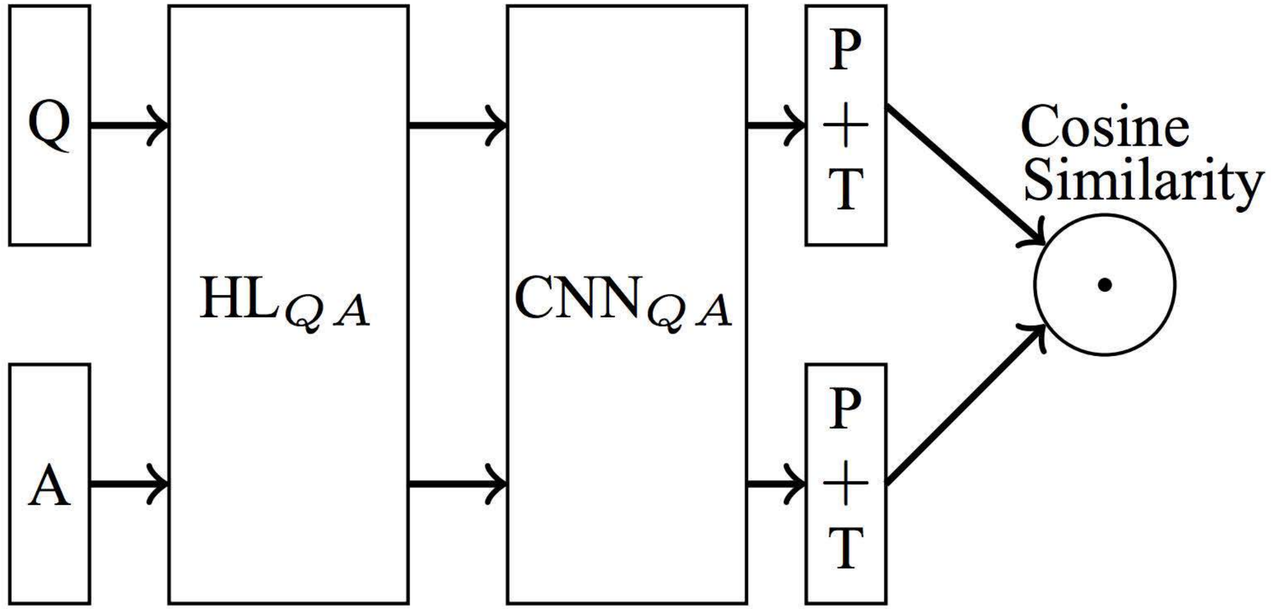
POS词性和句子主干有关

Word2vec的词向量本身具有语义上下文信息

算法	P准确率	R召回率	F-measture	效果提升
POS-IDF-词共现算法	0.624	0.345	0.445	25%

深度学习CNN

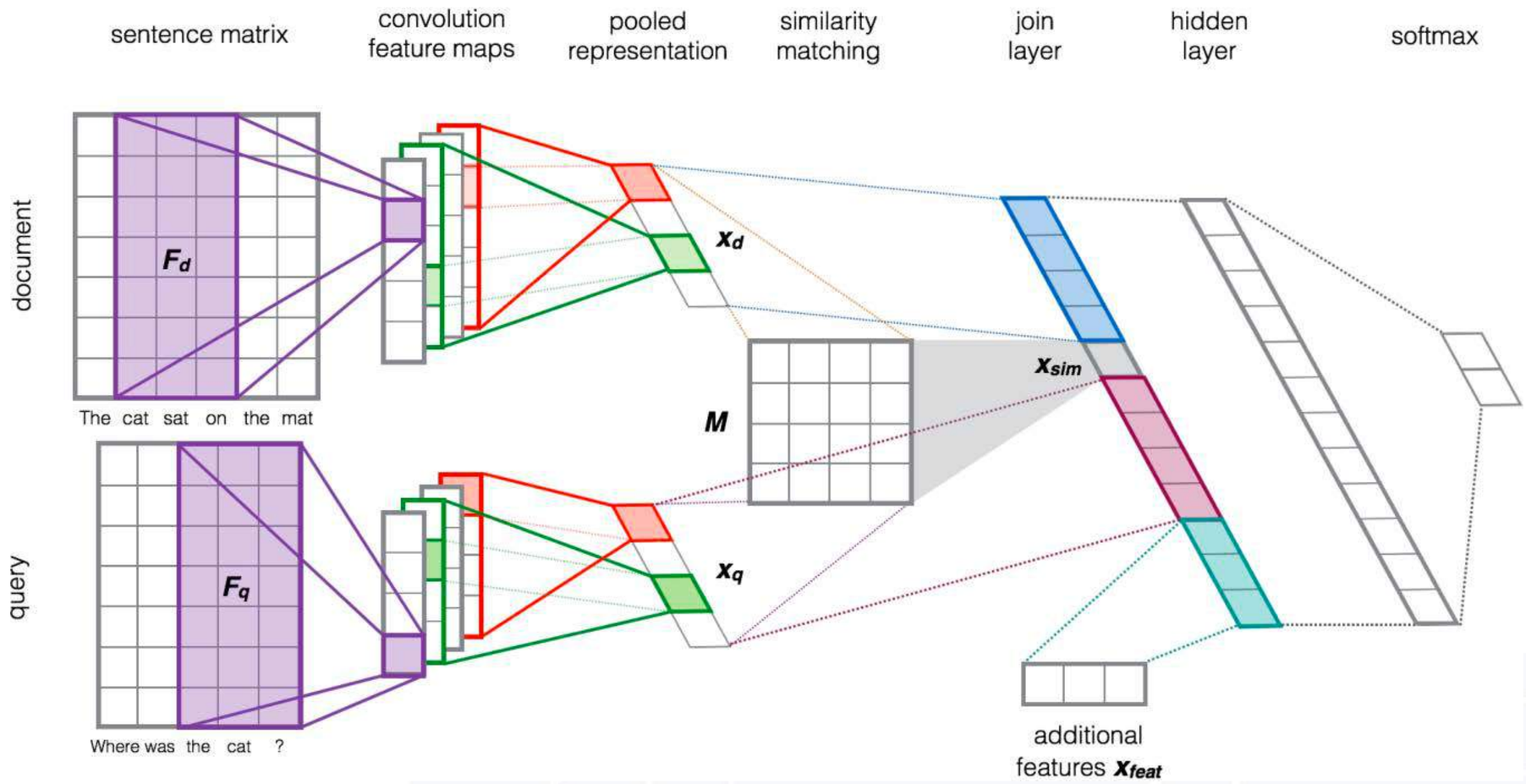
引入深度学习CNN模型，基于Tensorflow开发\GPU训练\评测\部署，QA数据做了拼接共用一个卷积池化的参数，然后求余弦相似度。



Applying Deep Learning to Answer Selection: A Study and An Open Task：https://arxiv.org/abs/1508.01585

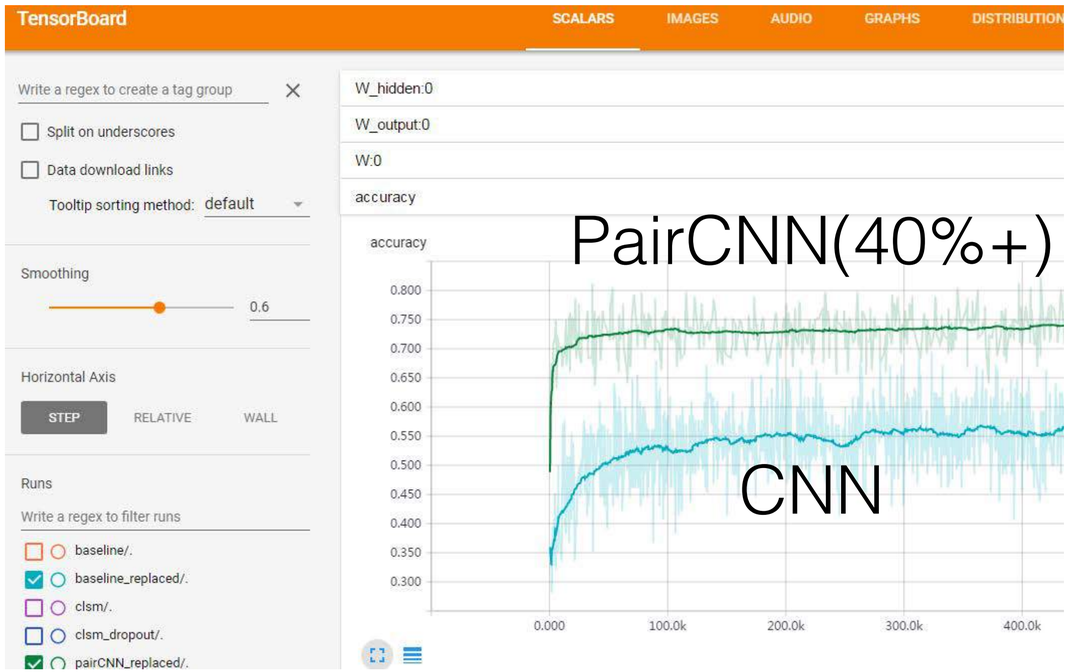
深度学习PairCNN

我们引入PairCNN改进模型对原有的CNN模型对比优化。



Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks：https://dl.acm.org/citation.cfm?id=2767738

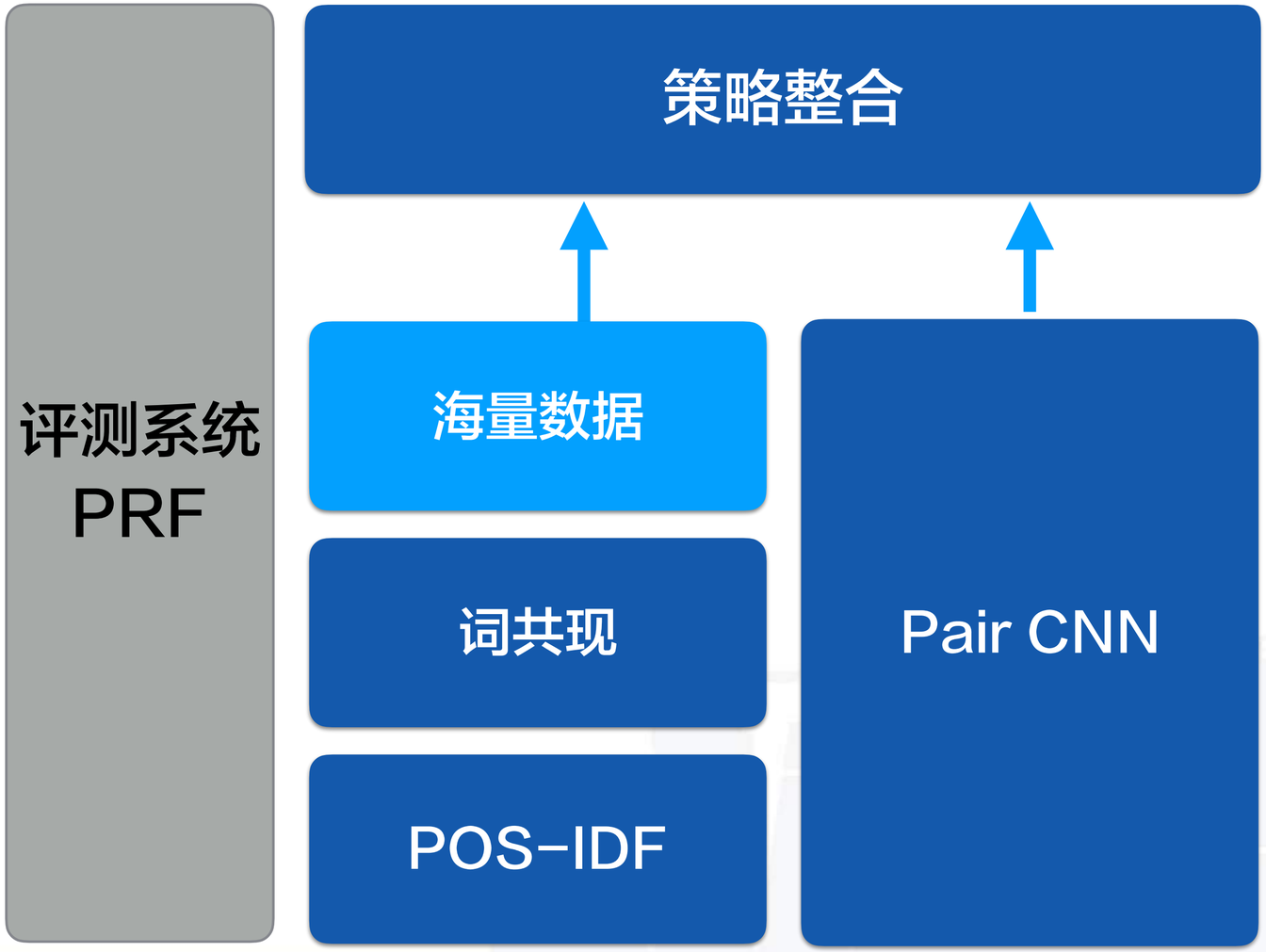
模型对比



模型	对比
CNN	1、共享所有卷积池化的参数，且卷积核数量较大，计算量大，对于Q和A的文本的区分度(diversity)无法分别建模； 2、Q和A的交互发生在最后MLP层中，表达相关性的能力有限
PairCNN	1、Q和A分别用不同的卷积池化参数进行建模，且卷积核数量小于原有模型，复杂度更低；2、在MLP层之前增加了Q与A相关性抽取的交互，从high-level的语义向量中进一步抽取了相关性特征，能更加丰富语义表征

算法	P准确率	R召回率	F-measture	效果提升
PairCNN算法	0.816	0.415	0.551	55.6%

技术特色——独家配方Rank模型



1、各自评测
2、综合配比

技术特色——独家配方Rank模型

综合词共现 / IDF词向量 / PairCNN取得比baseline在F值上约70%的提升。

F=2*P*R/(P+R)

算法	P准确率	R召回率	F-measture	效果提升
词共现算法	0.532	0.265	**0.354**	-----
POS-IDF-词共现算法	0.624	0.345	0.445	25%
PairCNN	0.816	0.415	0.551	55.6%
词共现+idf-word2vec	0.809	0.431	0.563	59%
PairCNN+词共现	0.812	0.448	0.577	63%
三种算法结合起来	**0.804**	**0.479**	**0.601**	**69.77%**

技术特色——文本情感分析

- 基于UGC中文和Emoji表情label混杂得到训练数据
- 利于深度学习LSTM模型进行文本情感分类
- 目前6分类的成功率大于80%，业界第一梯队水平标准了
- 结合宠物的情感分析商业价值也开始显现(KFC &宠物)

Agenda

1 聊天机器人

2 **腾讯云小微**

3 小Q机器人

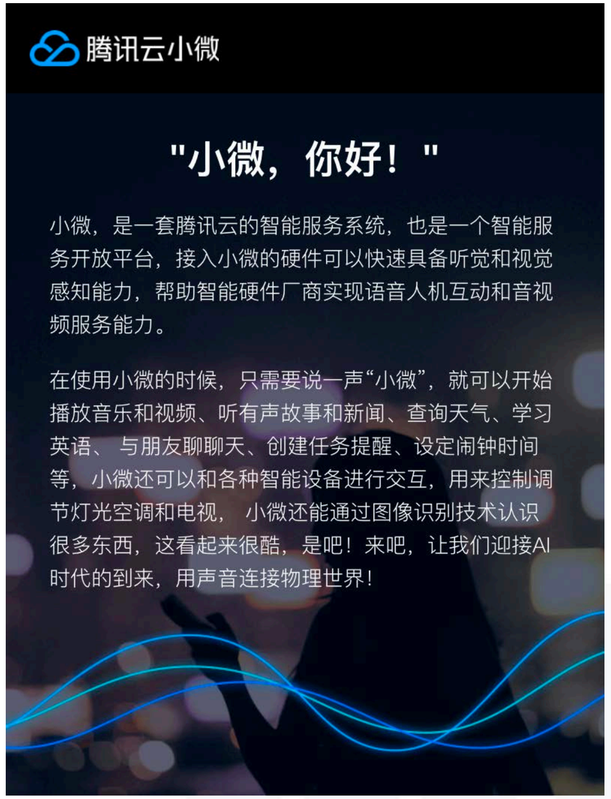
腾讯云小微

产品定位

小Q机器人

整合+开放+赋能

能力展示

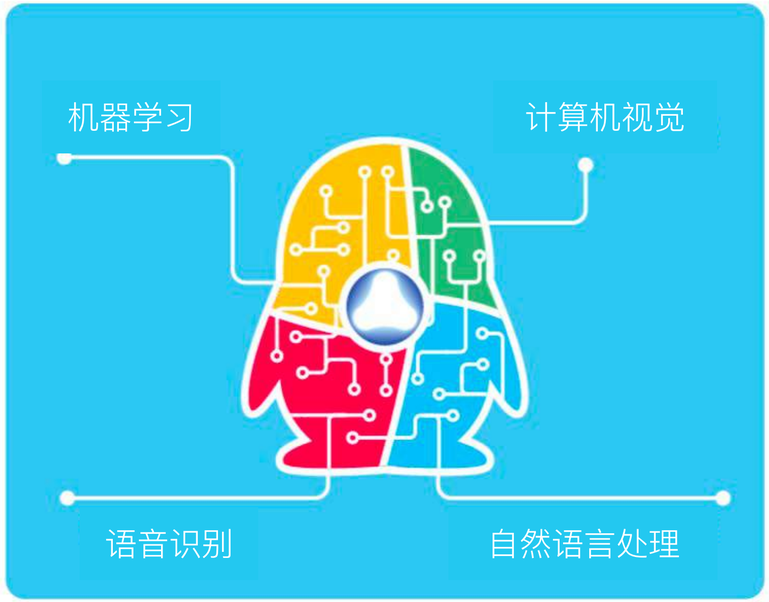


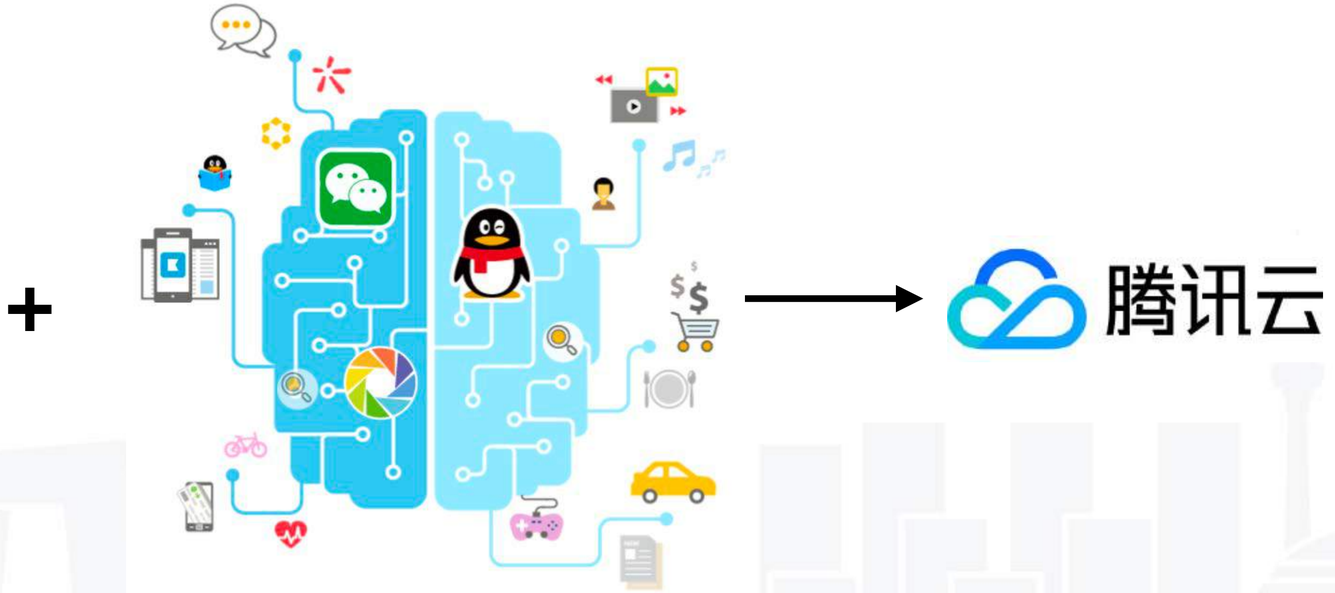
云服务
(腾讯云小微)

- 人工智能探索性产品
- 拥有多项AI能力的智能机器人
- 开放成为“腾讯云小微”云服务
- “腾讯云小微”的行业标杆产品

公司战略

人工智能，令人振奋的长远投资。
整合人工智能基础能力，赋能生态系统合作伙伴。——腾讯2017年Q2财报





能力整合

整合腾讯内多项AI能力。

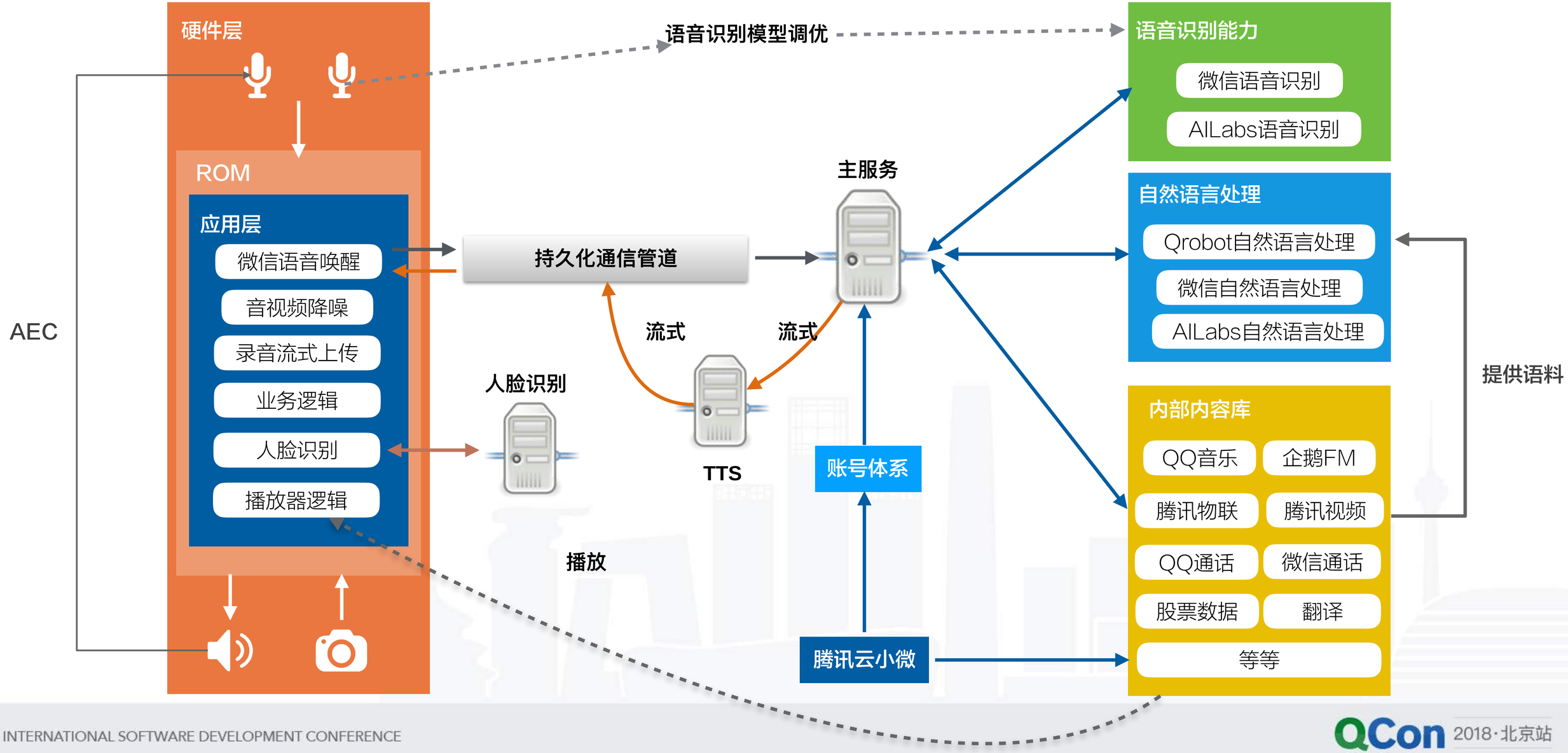
微信
AILabs
音视频实验室

Qrobot
微信
AILabs

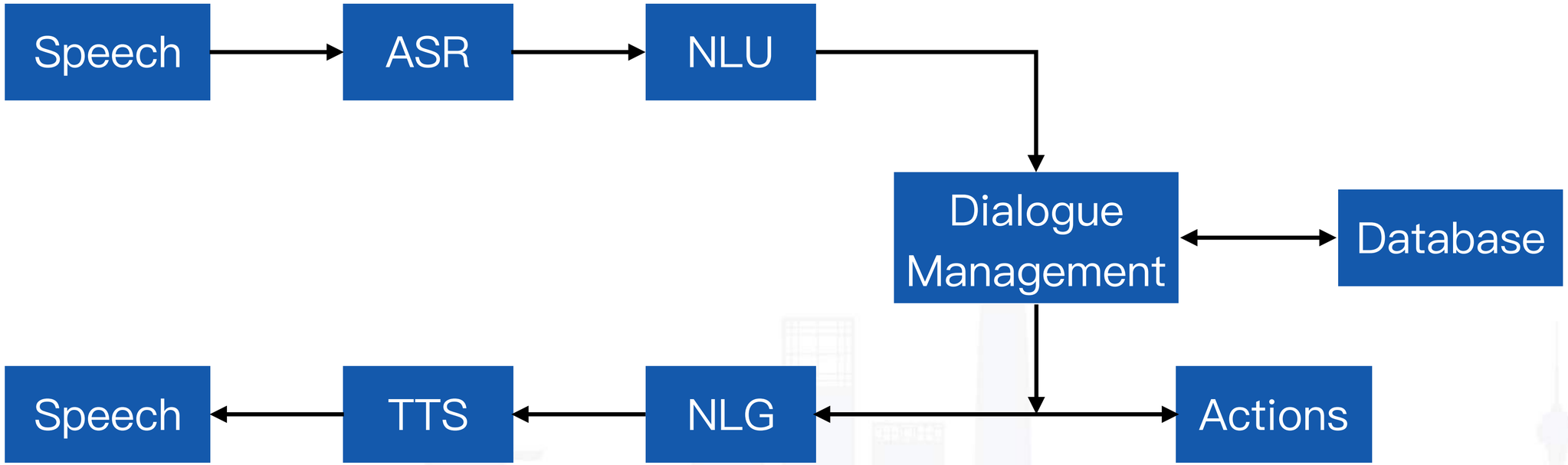
腾讯物联
QQ音乐
企鹅FM
腾讯视频
小企鹅乐园等等

优图实验室
天天P图

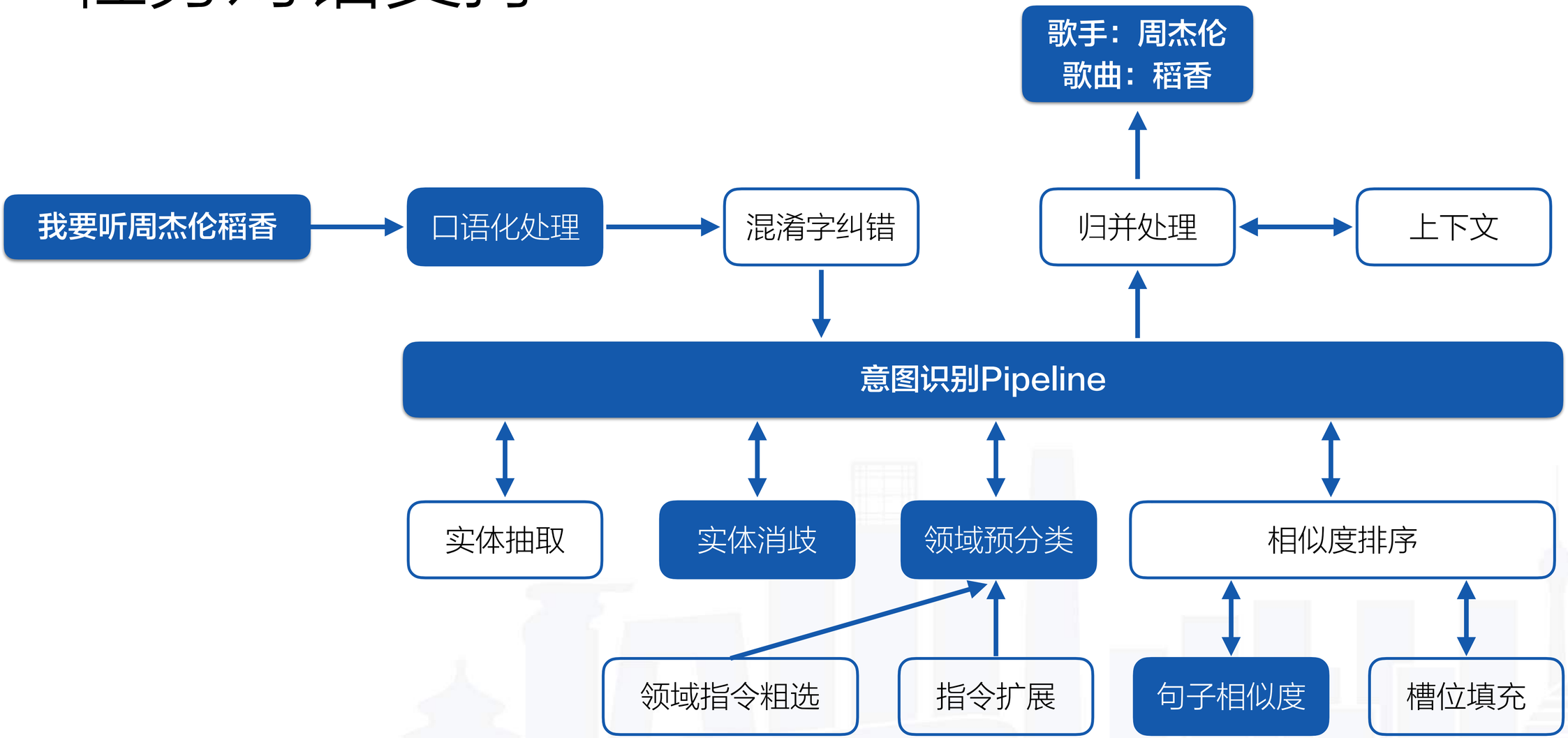
系统架构


硬件层
语音识别模型调优
语音识别能力
微信语音识别
AILabs语音识别
ROM
应用层
微信语音唤醒
音视频降噪
录音流式上传
业务逻辑
人脸识别
播放器逻辑
主服务
持久化通信管道
流式
流式
人脸识别
TTS
账号体系
腾讯云小微
自然语言处理
Qrobot自然语言处理
微信自然语言处理
AILabs自然语言处理
内部内容库
QQ音乐
企鹅FM
腾讯物联
腾讯视频
QQ通话
微信通话
股票数据
翻译
等等
AEC
播放
提供语料

任务对话支持

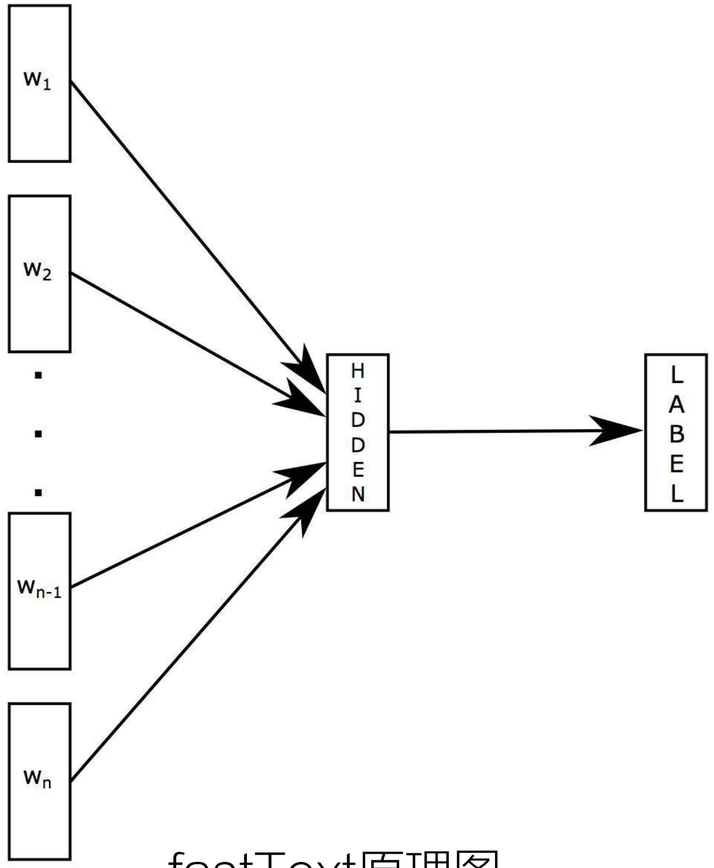

Speech
ASR
NLU
Dialogue Management
Database
Speech
TTS
NLG
Actions

任务对话支持


歌手：周杰伦
歌曲：稻香
我要听周杰伦稻香
口语化处理
混淆字纠错
归并处理
上下文
意图识别Pipeline
实体抽取
实体消歧
领域预分类
相似度排序
领域指令粗选
指令扩展
句子相似度
槽位填充

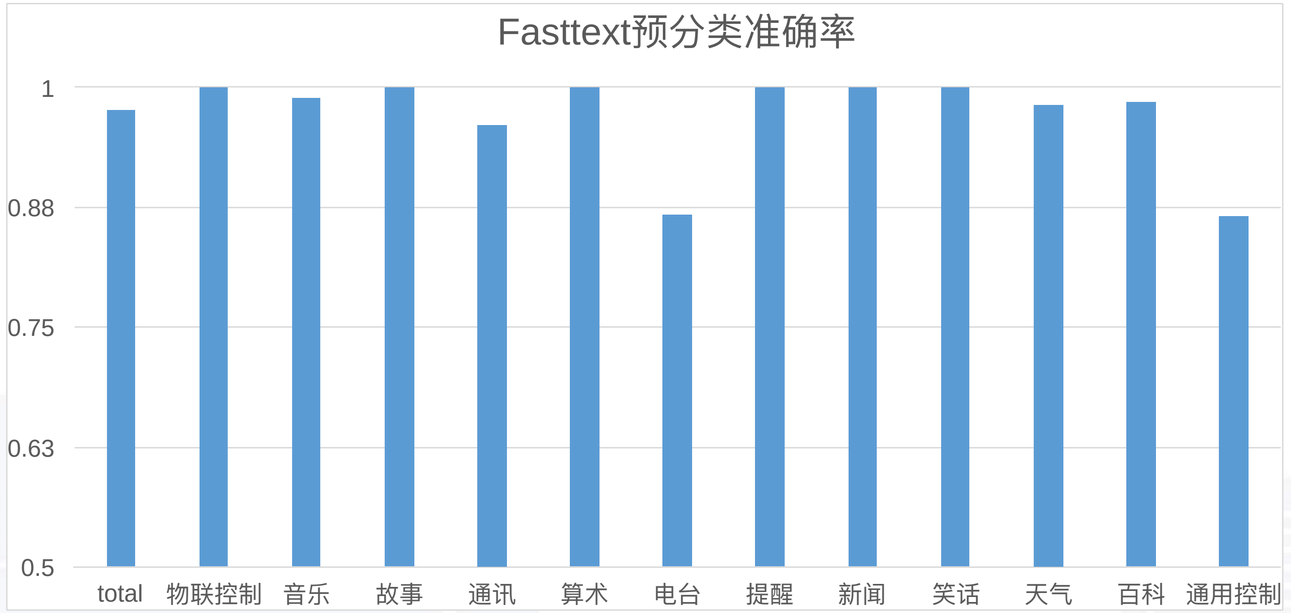
技术特色——领域预分类

提高响应速率策略：fastText预分类模型。



fastText原理图

优点是它是浅层神经网络，计算简单；
在top3的领域上选中目标准确率达到97.6%。



平均准确率：0.976

技术特色——句子相似度

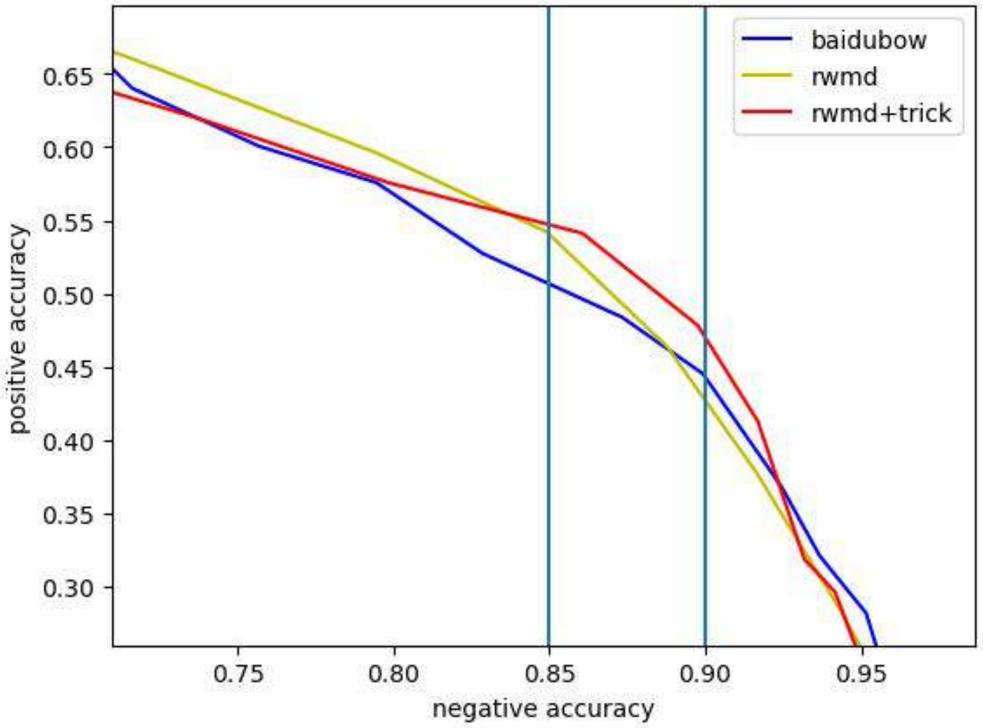
提高正确率策略：改进语义相似度算法

主流语义相似度算法：CNN, wmd

Qrobot语义相似度算法：rwmd+**调整策略**

1.由于停用词影响正确率（你、我、他等），这里设置停用词去掉后剩余词多于3个（谓宾短语），则去掉停用词。否则保留停用词。

2.另外对于slot的实体（歌曲、歌手等）给予较低权重。

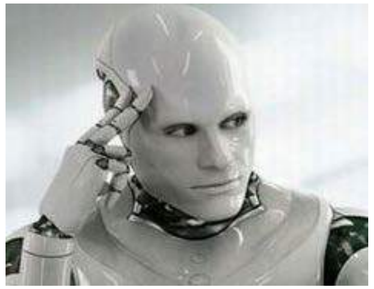


添加调整策略后正负样本准确率曲线对比

结论：

采用rwmd+调整策略方法后系统语义相似度的正确率拥有明显提升。

技术特色——实体消歧

听张学友 心如刀割 →  → 张学友 心如刀割

看深南大道 堵车 → → 看深南大道 堵车 (无实体)

基于29个句子/词性特征，实体消歧采用**GBDT+Boosting** 准确率达到**95%**，多个模型多轮提升效果如下图：

使用十折交叉验证的方法，使用 6 种常用分类模型对全量数据进行了拟合和效果验证，

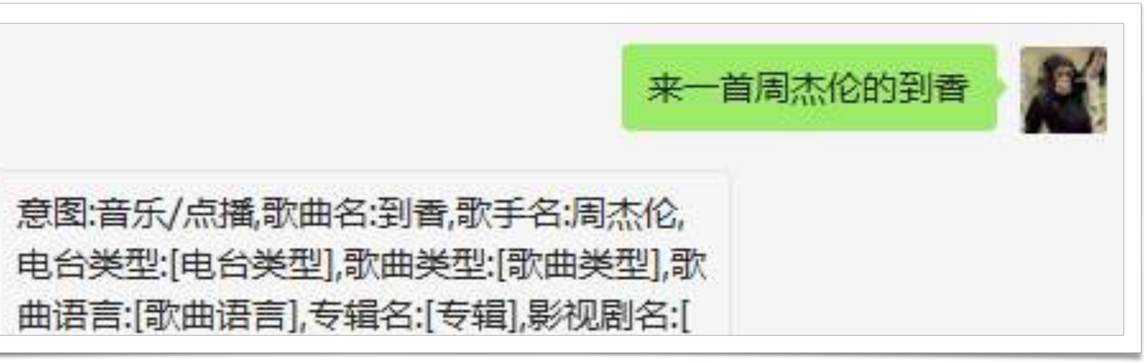
	Native Bayes	LR	SVM	决策树	GBDT	GBDT+Boost
第一轮特征	0.661	0.850	0.878	0.852	0.906	0.893
第二轮特征	0.716	0.858	0.882	0.862	0.908	0.898
第三轮特征	0.676	0.867	0.884	0.877	0.912	0.908

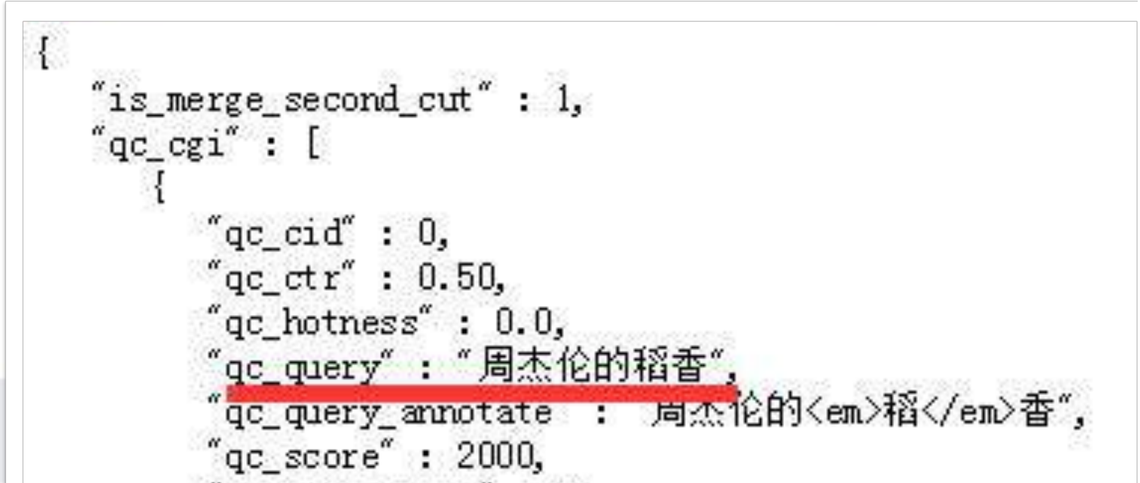
按照 7:3 的比例进行训练测试集划分之后，各个模型的表现如下表所示：

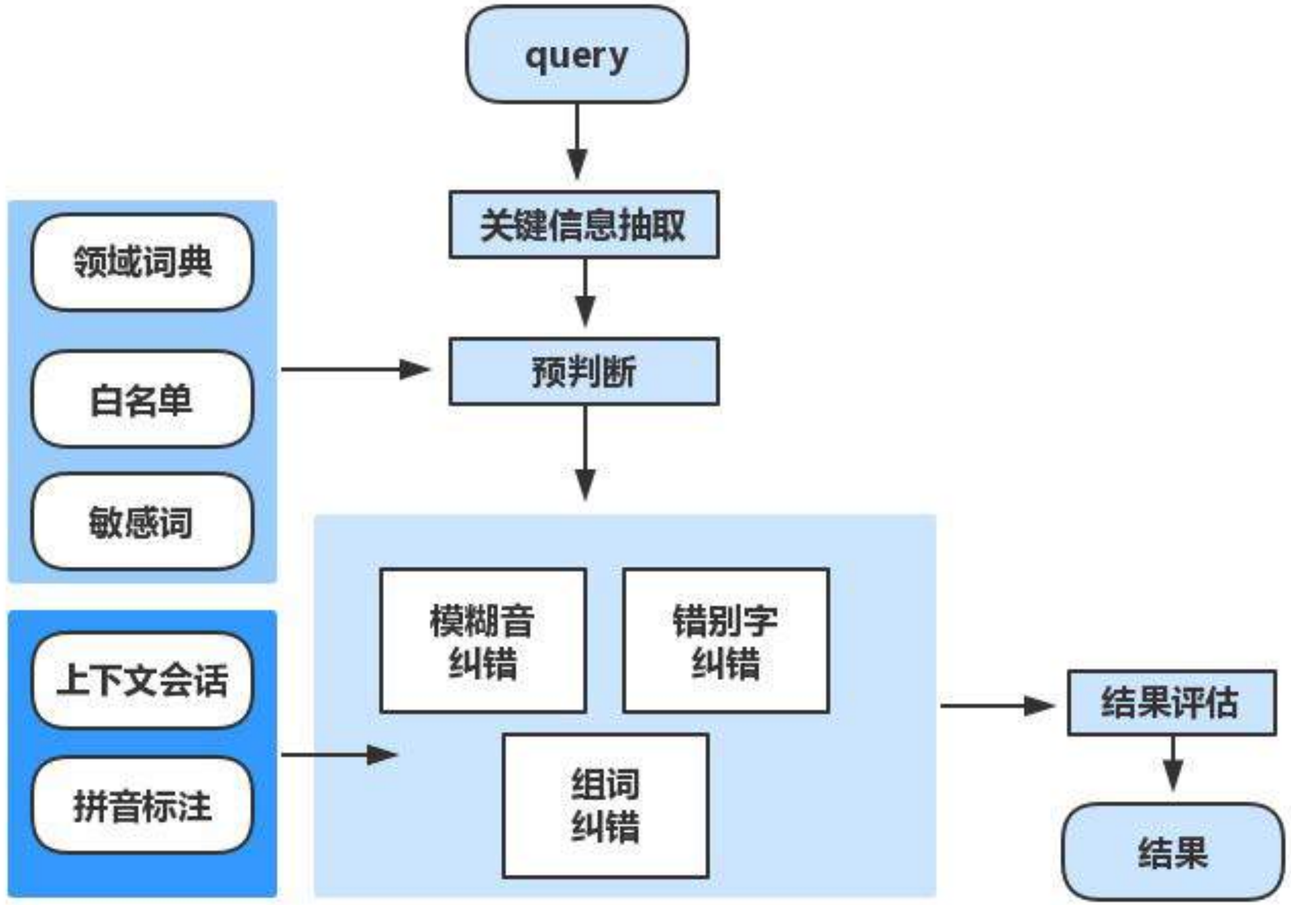
	Native Bayes	LR	SVM	决策树	GBDT	GBDT+Boost
第三轮特征	0.60	0.87	0.89	0.88	0.91	0.95

技术特色——口语化处理

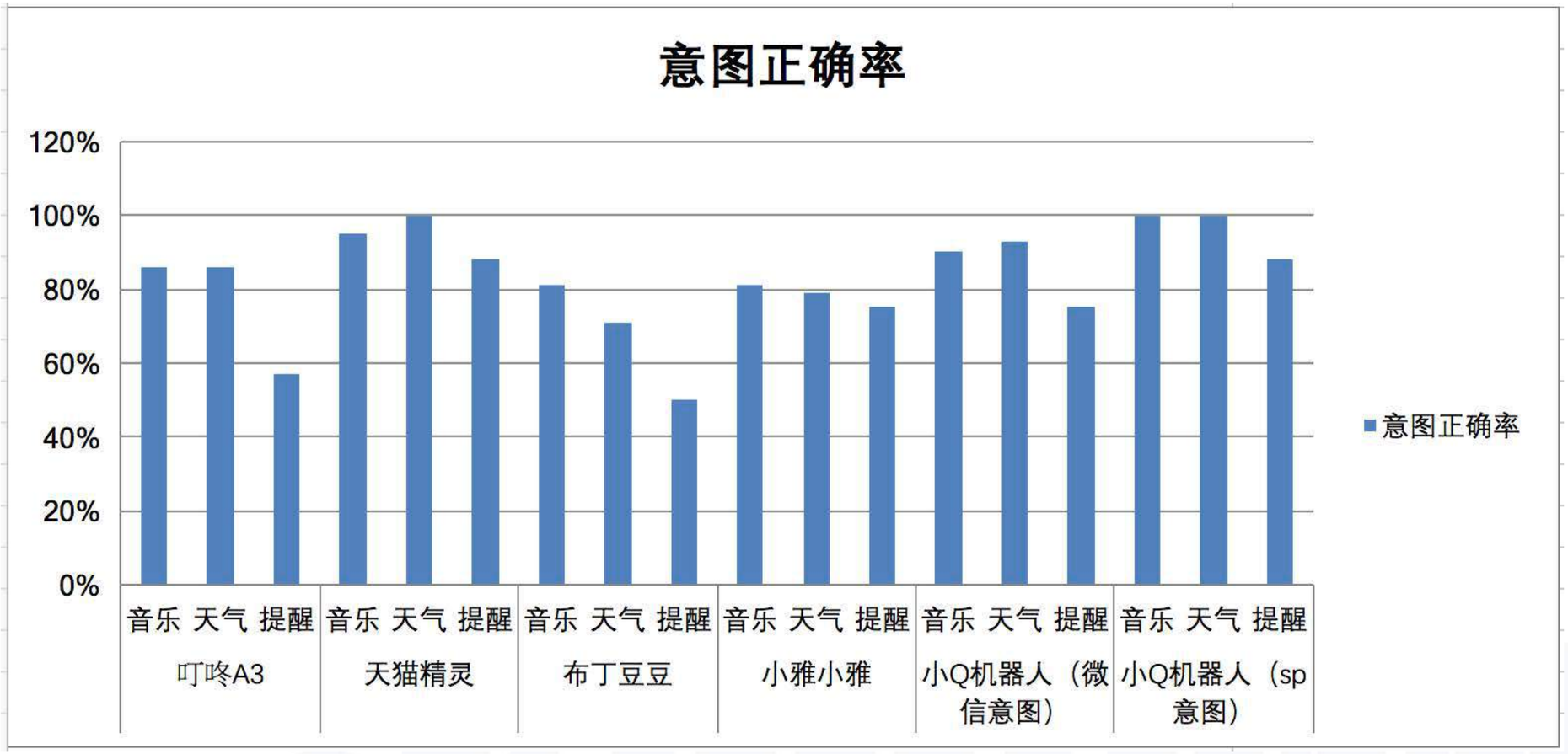
提高容错能力策略：Query Correction(QC)实现纠错







技术成果


意图正确率
120%
100%
80%
60%
40%
20%
0%
音乐 天气 提醒
叮咚A3
音乐 天气 提醒
天猫精灵
音乐 天气 提醒
布丁豆豆
音乐 天气 提醒
小雅小雅
音乐 天气 提醒
小Q机器人（微信意图）
音乐 天气 提醒
小Q机器人（sp意图）
意图正确率

Agenda

1 聊天机器人

2 腾讯云小微

3 小Q机器人

小Q机器人

面临的挑战

项目挑战

- 团队缺乏硬件基因，硬件如何选型，硬件研发如何开展?
- 硬件项目周期长，如何做到快速验证，快速试错，快速决策
- 消费级产品，软硬件调优，保障基础体验，保障质量

硬件拆解

多阵列麦克风

平板方案

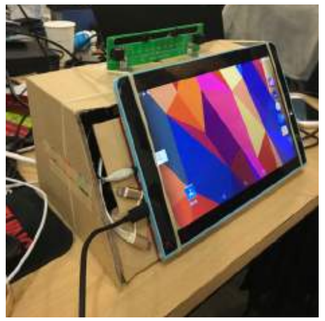
扬声器(音响)

落地方案

2017.03

2017.10

成熟硬件产品上快速打磨软件



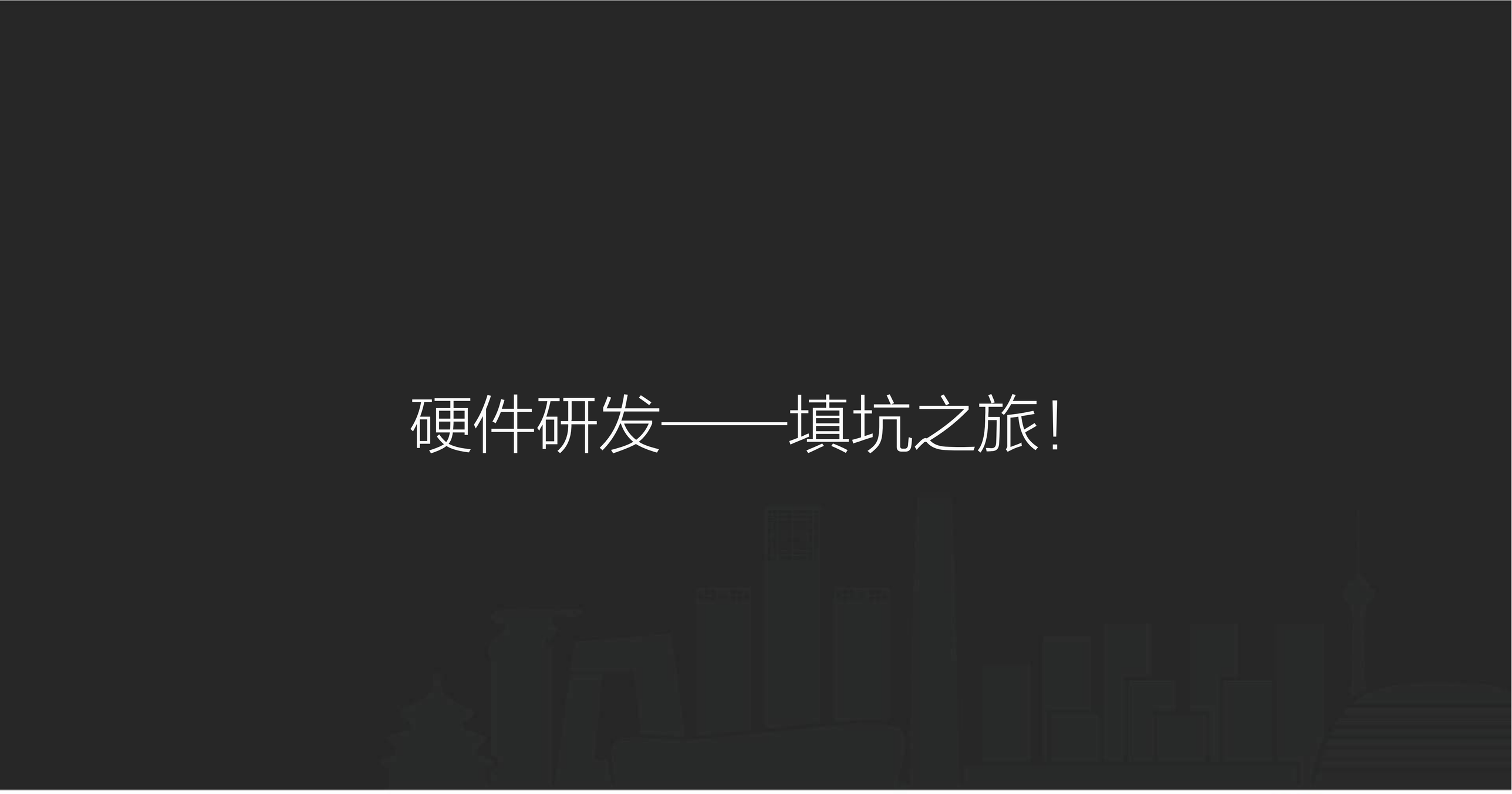
硬件研发，原型机快速完善软件

软硬结合优化

硬件问题


响应速度慢
MCU无法OTA升级
整机过热
有色差
低亮度下曝光不足
噪音大
唤醒率低
摄像头效果差
花屏
叠影
麦克风偶现失效
开机音量过大
电磁干扰
按键通信问题
无法充电
突然掉电
频繁唤醒
USB无法连接
交互奇怪

硬件研发——填坑之旅！



响应速度优化

第一个版本情况：

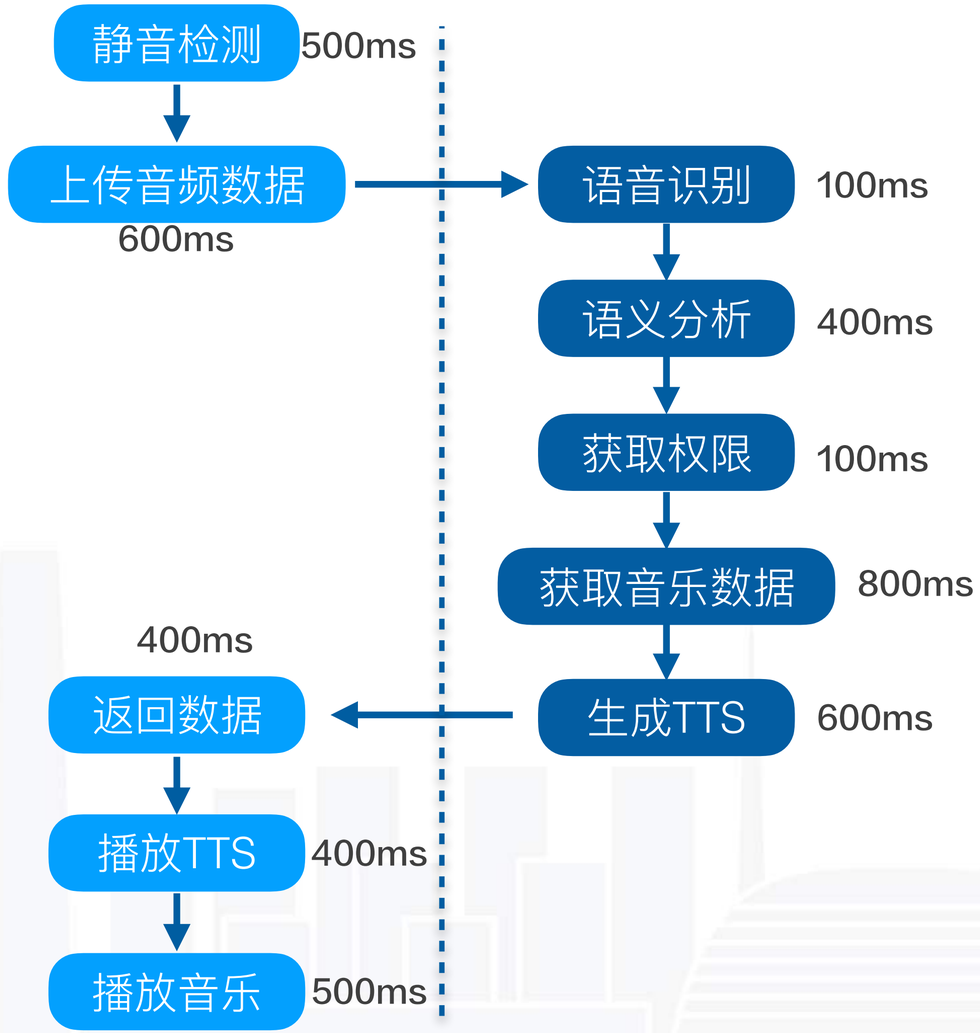
1、总耗时达4.4秒！
2、大部分步骤都必须串行；
3、每个流程都在不同的部门，有些功能还是刚走出实验室；
4、TTS耗时过长，需要1秒多；
5、静音检测每次都要消耗500ms；
6、后台耗时过长，接近2秒。

常用优化措施：

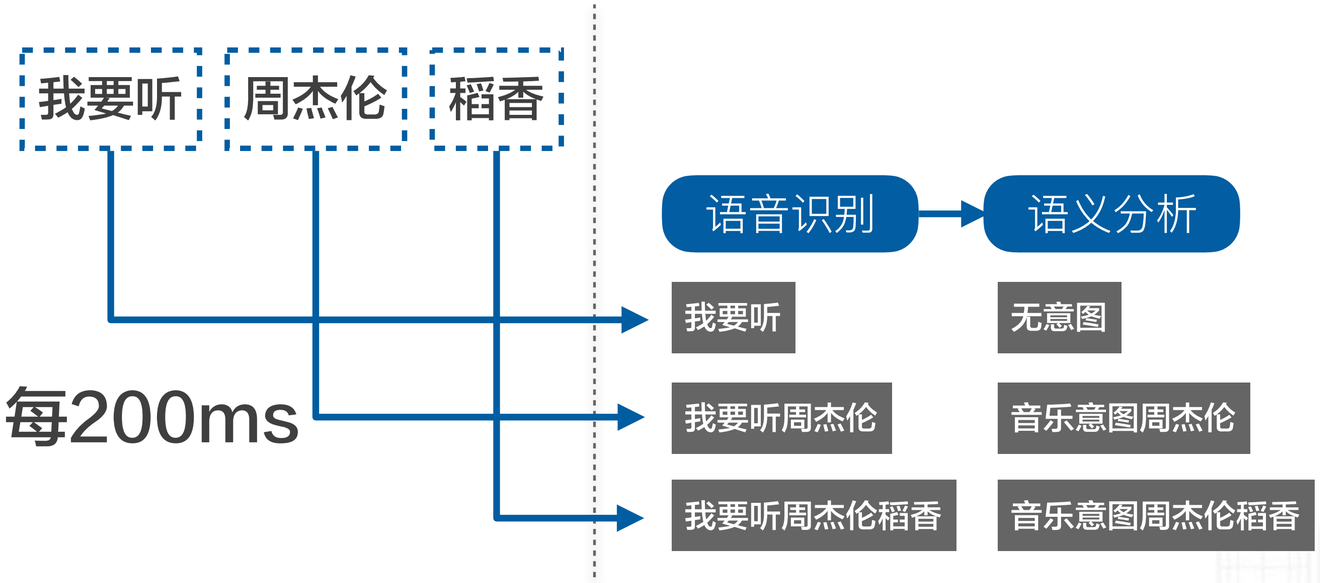
1、后台服务就近部署，并分离研发和线上环境，减少不必要逻辑；
2、优化各个步骤耗时，比如音乐从800ms优化到400ms；
3、一些常用指令做数据缓存。

目标是秒开，追赶Amazon Echo的应答速度。

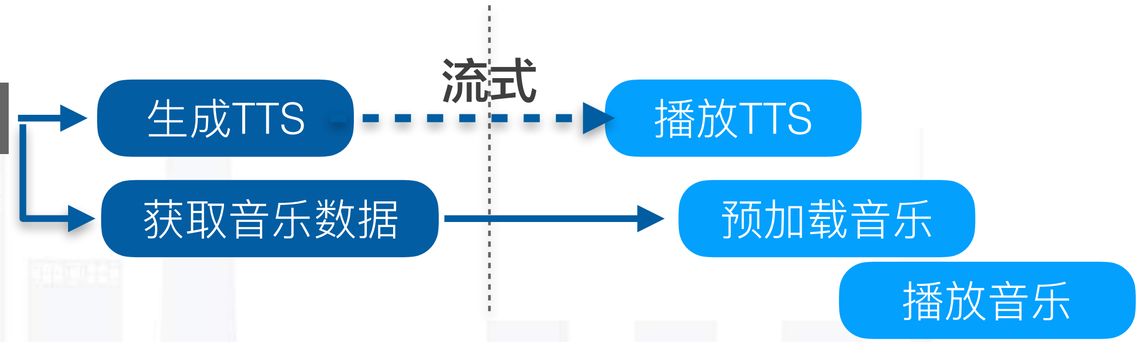
耗时仍然无法满足要求。



响应速度优化—全流程流式



3、TTS不用等待音乐查询结果，节省约400ms；
4、TTS改为流式生成，流式传输，流式播放，节省约800ms；
5、播放TTS时，预加载音乐内容，TTS播放结束，音乐可以立即播放，节省约200ms。

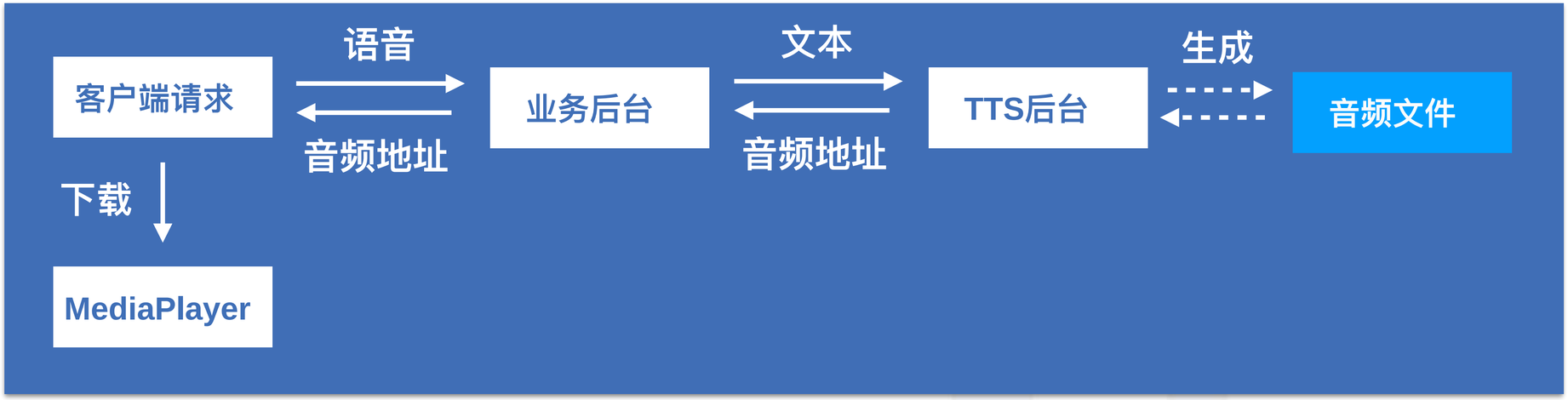


1、语音识别和语义分析改为流式计算，检测到静音的同时，就能拿到语义的结果，节省耗时1000ms以上；
2、静音检测VAD逻辑迁移到后台，算法调整更灵活，节省约200ms。

响应速度优化——TTS流式优化

传统TTS模式：

一次性生成音频文件，返回地址客户端下载播放。



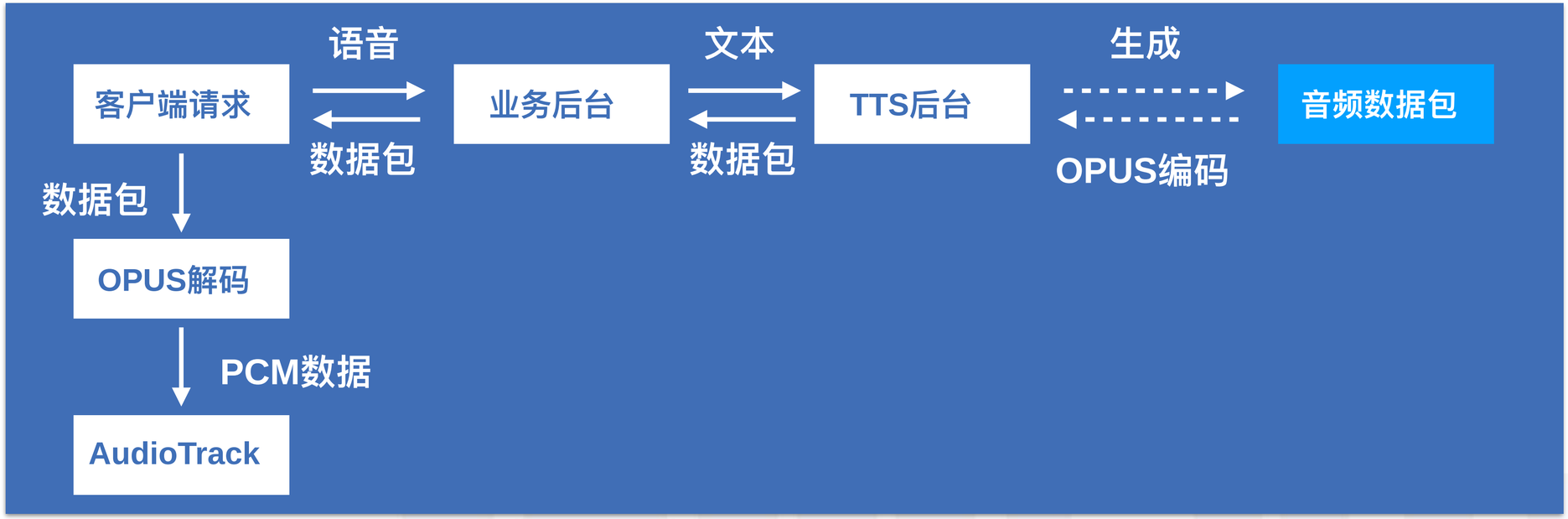
痛点：耗时！（在百科类长内容上非常明显）

1、TTS后台生成完整音频文件时间
2、客户端拿到音频地址后时间到可以播放的时间

响应速度优化——TTS流式优化

流式TTS模式：

根据文本实时生成语音包数据，借助TCP长链接，OPUS将语音包编码后，有序送往客户端解码播放，OPUS可将语音包从2880b减至300b。



与传统模式对比优势：

1、流式TTS的播放时间不依赖于文本的长度，在长文本和短文本上时间无差距

2、客户端无需要下载，长链接会以PUSH的形式有序传递给客户端解码播放