CODE A BETTER LIFE
一 行 代 码　亿 万 生 活

AI语音产品测试

演讲人：于海生

美团

于海生

美团技术专家
3年开发经验，9年测试经验。
先后经历过音视频中间件、ERP中间件、搜索、自然语言交互等领域测试

智能语音应用场景有哪些?

智能语音产品如何测试?

主题

- **AI简介**
- 智能语音技术应用场景
- 如何测试语音产品并评定质量
- QA

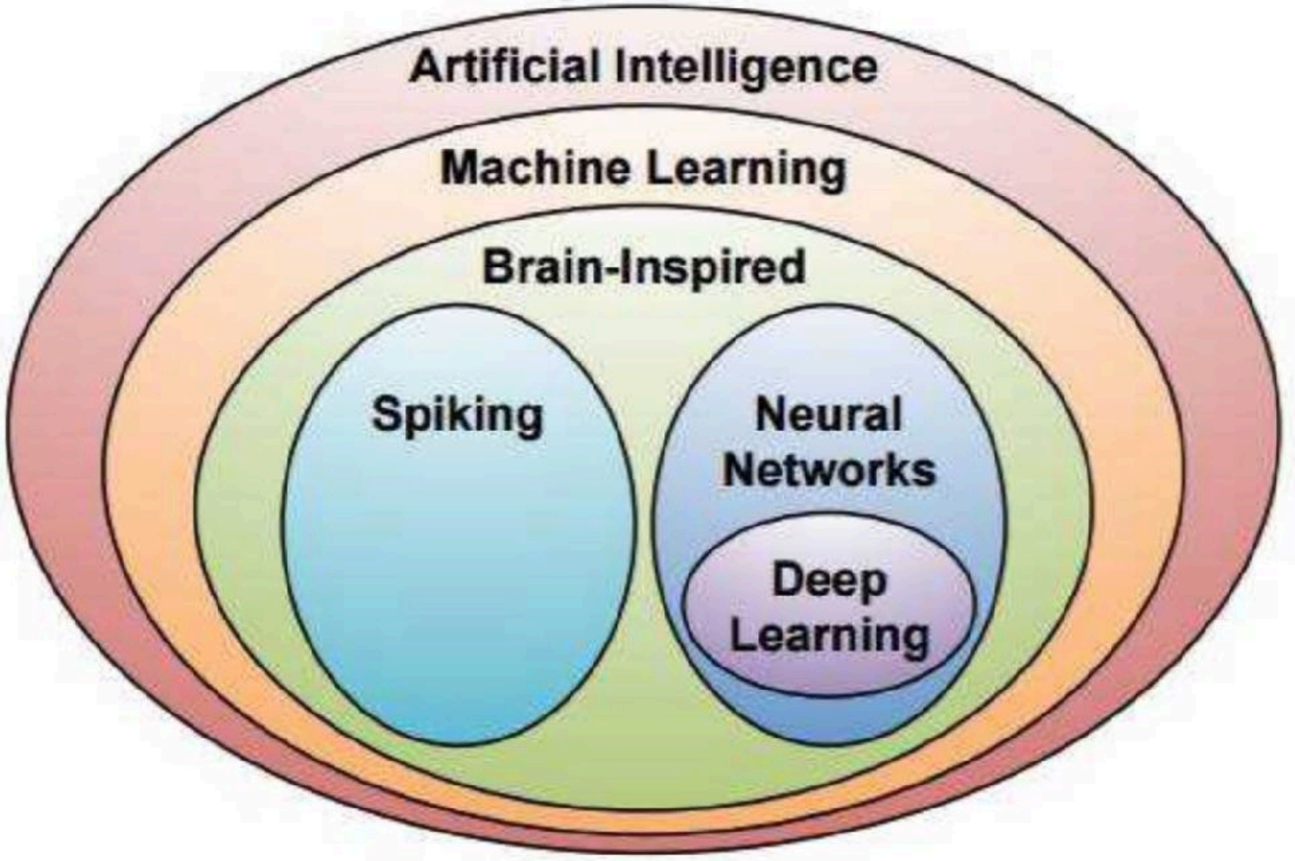
techopedia

Artificial intelligence (AI) is an area of computer science that emphasizes the creation of intelligent machines that work and react like humans.

wikipedia

Artificial intelligence (**AI**), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and other animals.

AI vs ML vs NN vs DL



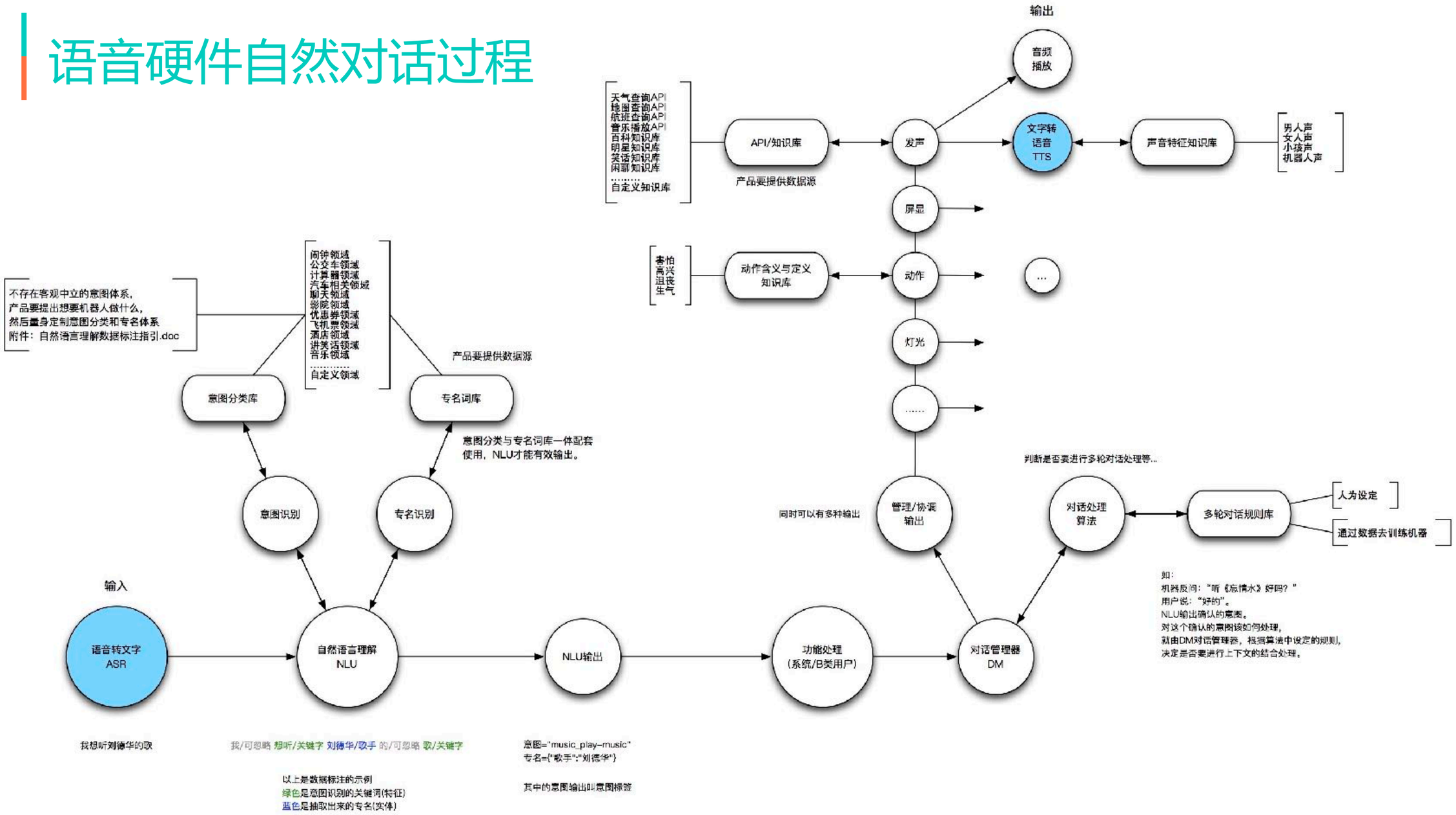
- AI简介
- **智能语音技术应用场景**
- 如何测试语音产品并评定质量
- QA


4G 2.7K/s
18:05
HD 4G 47%
我是小V,有什么可以帮您的吗？
哈尔滨天气怎么样
哈尔滨今天多云，最低温度15℃，最高温度25℃。
-22°C
15-25°C
多云
哈尔滨
周三
15~22°C
周四
15~25°C
周五
15~25°C
周六
11~21°C

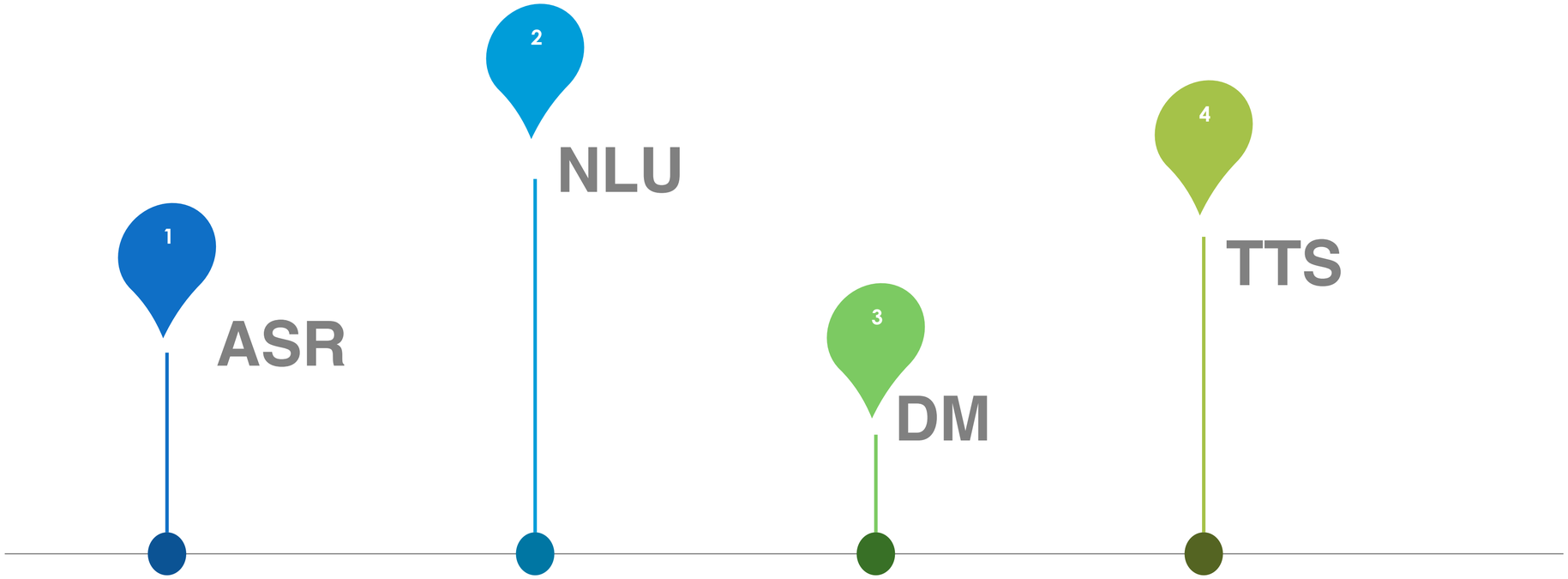

What can I help you with?

- AI简介
- 智能语音技术应用场景
- **如何测试语音产品并评定质量**
- QA

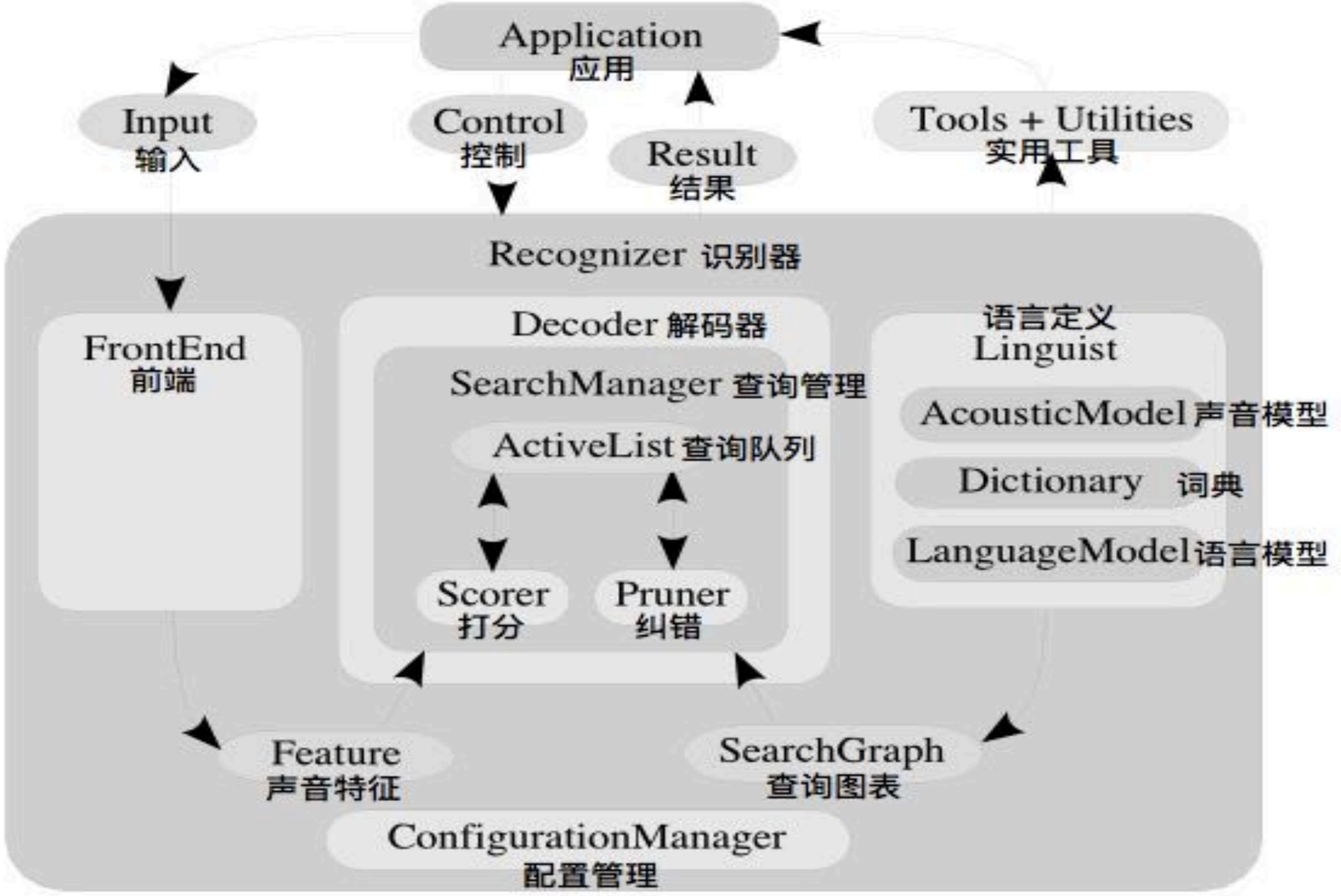
语音硬件自然对话过程


输出
音频播放
天气查询API
地图查询API
航班查询API
音乐播放API
百科知识库
明星知识库
笑话知识库
闲聊知识库
……
自定义知识库
API/知识库
产品要提供数据源
发声
文字转语音TTS
声音特征知识库
男人声
女人声
小孩声
机器人声
屏显
害怕
高兴
沮丧
生气
动作含义与定义知识库
动作
…
灯光
……
不存在客观中立的意图体系，
产品要提出想要机器人做什么，
然后量身定制意图分类和专名体系
附件：自然语言理解数据标注指引.doc
闹钟领域
公交车领域
计算器领域
汽车相关领域
聊天领域
影院领域
优惠券领域
飞机票领域
酒店领域
讲笑话领域
音乐领域
……
自定义领域
意图分类库
专名词库
产品要提供数据源
意图分类与专名词库一体配套使用，NLU才能有效输出。
意图识别
专名识别
判断是否要进行多轮对话处理等...
同时可以有多种输出
管理/协调输出
对话处理算法
多轮对话规则库
人为设定
通过数据去训练机器
如：
机器反问："听《忘情水》好吗？"
用户说："好的"。
NLU输出确认的意图。
对这个确认的意图该如何处理，
就由DM对话管理器，根据算法中设定的规则，
决定是否要进行上下文的结合处理。
输入
语音转文字ASR
自然语言理解NLU
NLU输出
功能处理（系统/B类用户）
对话管理器DM
我想听刘德华的歌
我/可忽略 想听/关键字 刘德华/歌手 的/可忽略 歌/关键字
以上是数据标注的示例
绿色是意图识别的关键词(特征)
蓝色是抽取出来的专名(实体)
意图="music_play-music"
专名={"歌手":"刘德华"}
其中的意图输出叫意图标签

语音产品核心架构图



ASR语音识别处理过程



ASR
质量指标

A 识别率

- 客户体验角度

识别率：100-SER（句错率）

SER: 句子识别错误的个数/总的句子个数

- 工程角度

识别率：100-WER（word error rate，词错率）

分为男女、快慢、口音、数字、英文、中文。

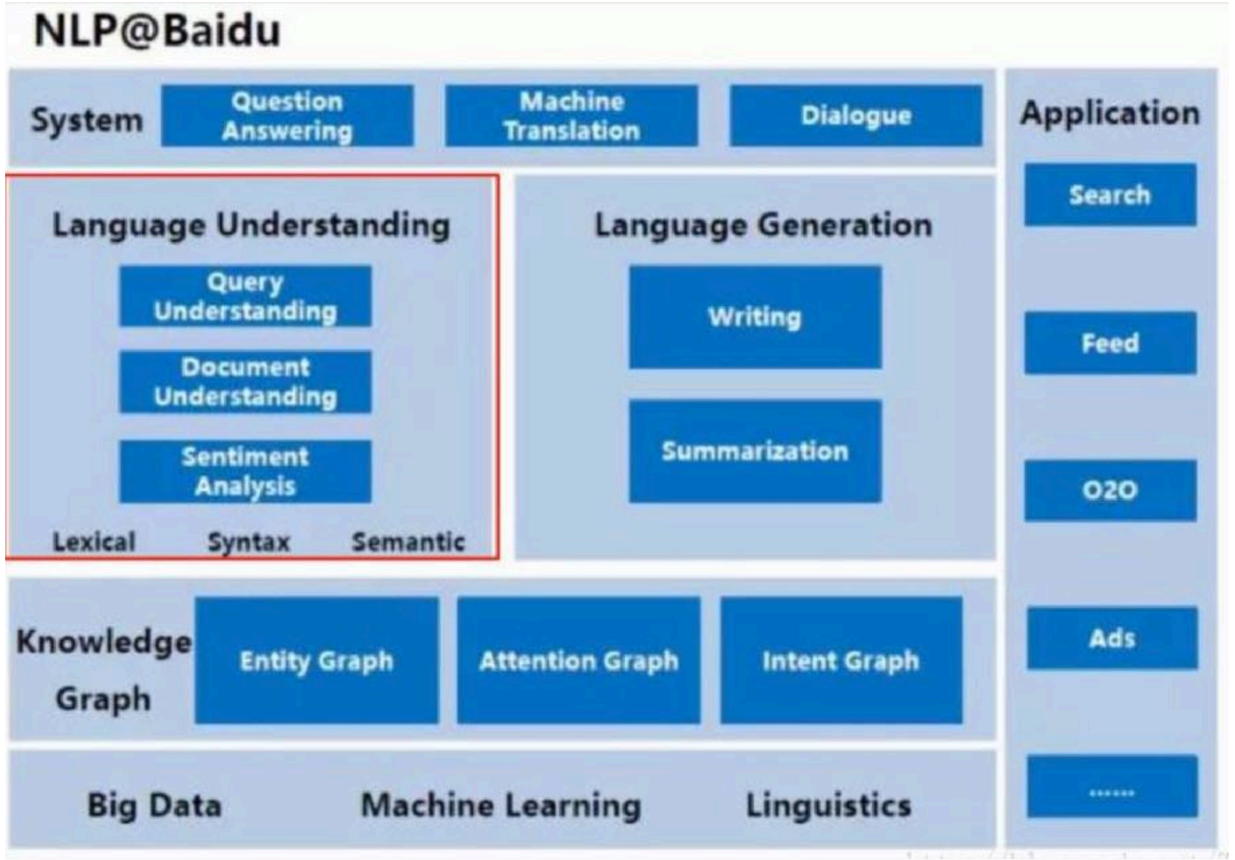
B 语音唤醒 VOICE TRIGGER，VT：

- 唤醒率
- 误唤醒率：没叫AI的时候，ta自己跳出来讲话的比率
- 唤醒词的音节长度：一般技术上要求，最少3个音节
- 唤醒响应的时间
- 功耗

NLU

自然语言理解
Natural Language Understanding, NLU

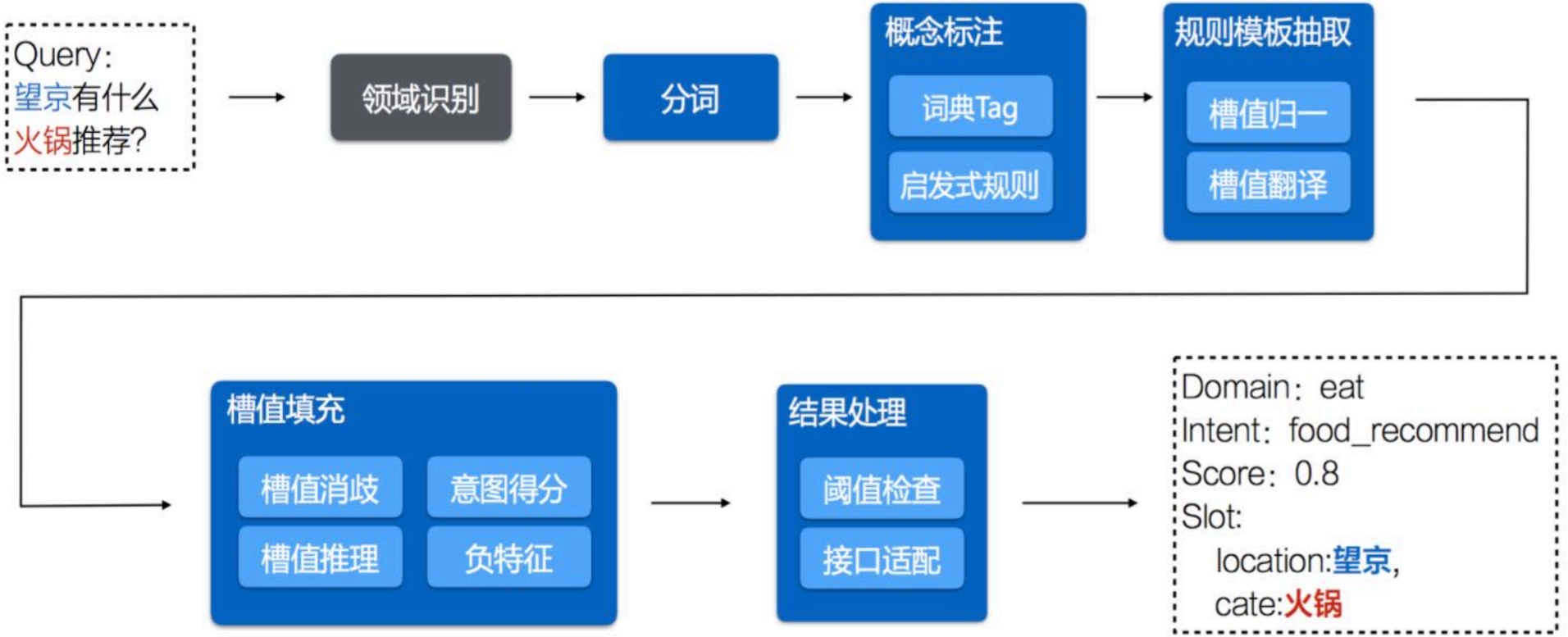
- 句子检测
- 分词
- 词性标注
- 句法分析
- 文本分类/聚类
- 信息抽取/自动摘要等



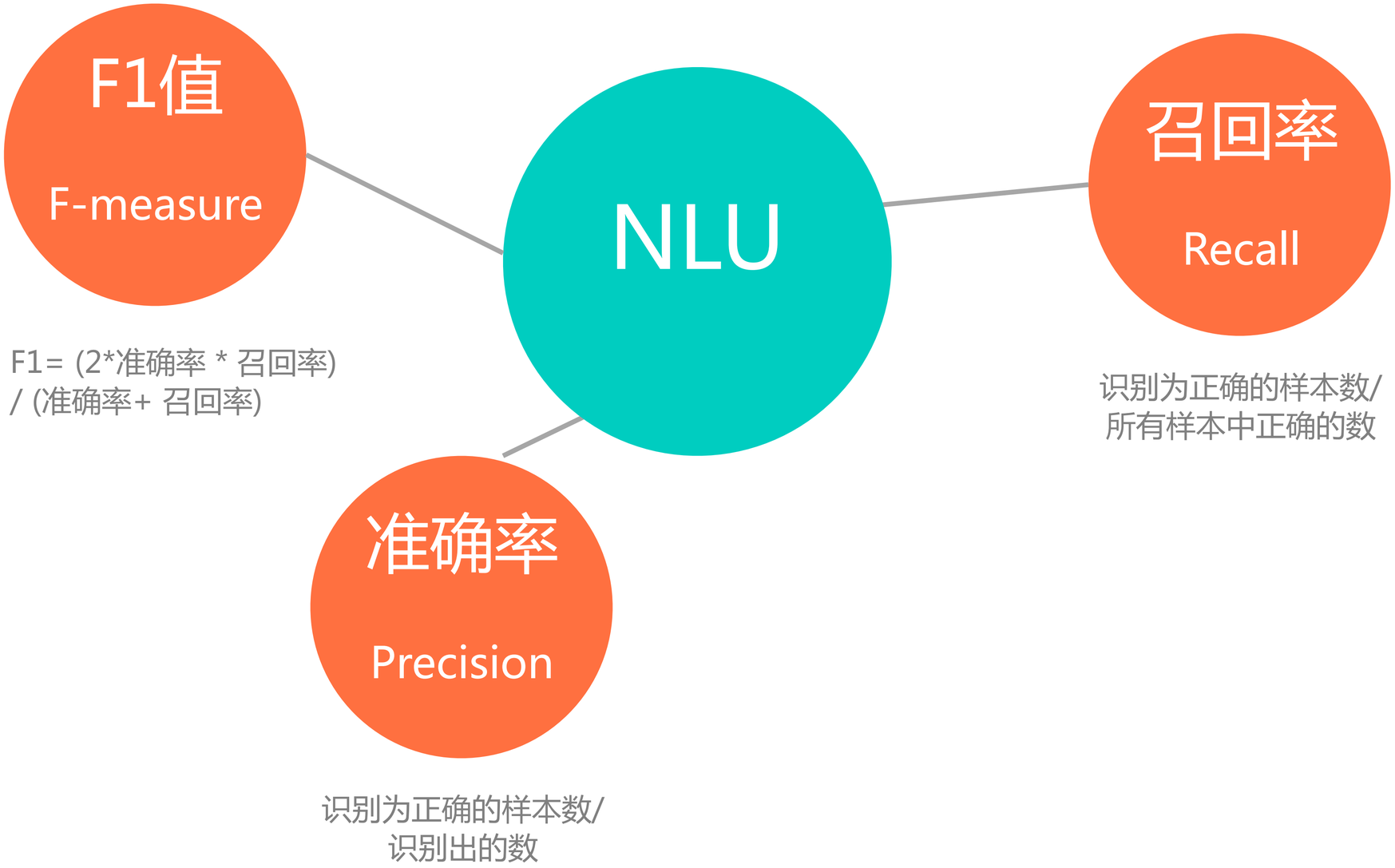
自然语言处理
Natural Language Processing，NLP

- 是自然语言和AI交集的学科，包括语音处理相关的所有技术

NLU处理过程

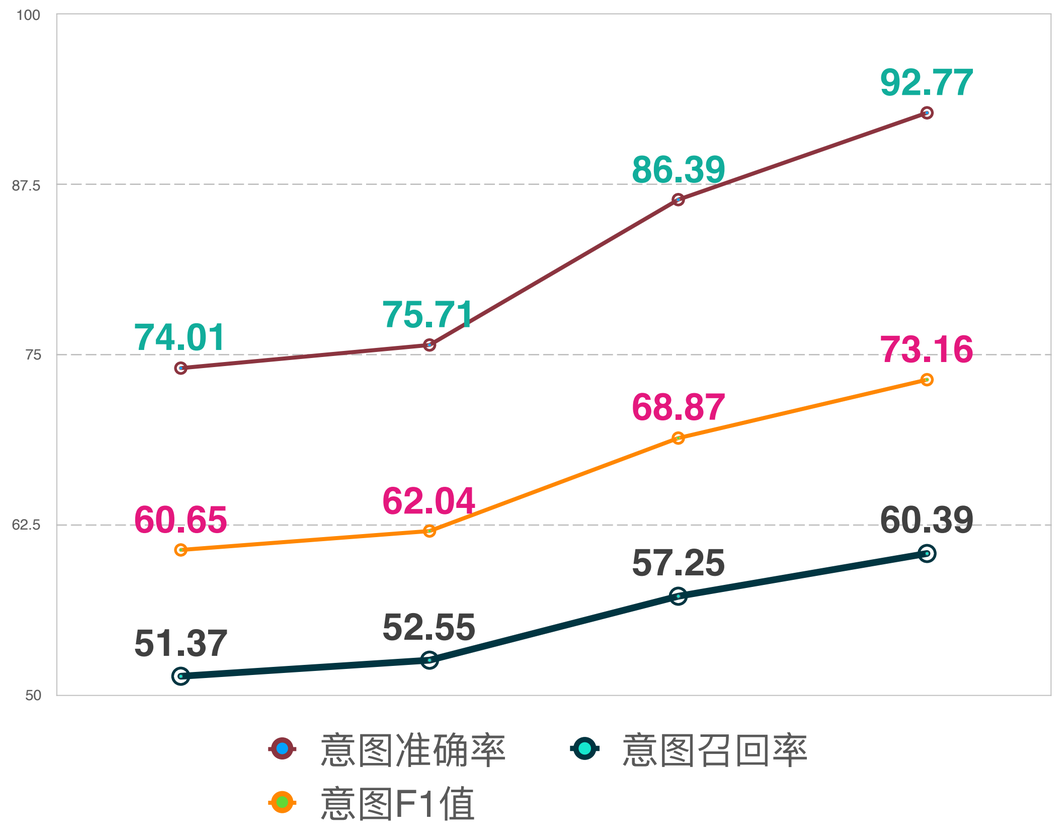

Query:
望京有什么
火锅推荐?
领域识别
分词
概念标注
词典Tag
启发式规则
规则模板抽取
槽值归一
槽值翻译
槽值填充
槽值消歧
意图得分
槽值推理
负特征
结果处理
阈值检查
接口适配
Domain：eat
Intent：food_recommend
Score：0.8
Slot:
location:望京,
cate:火锅

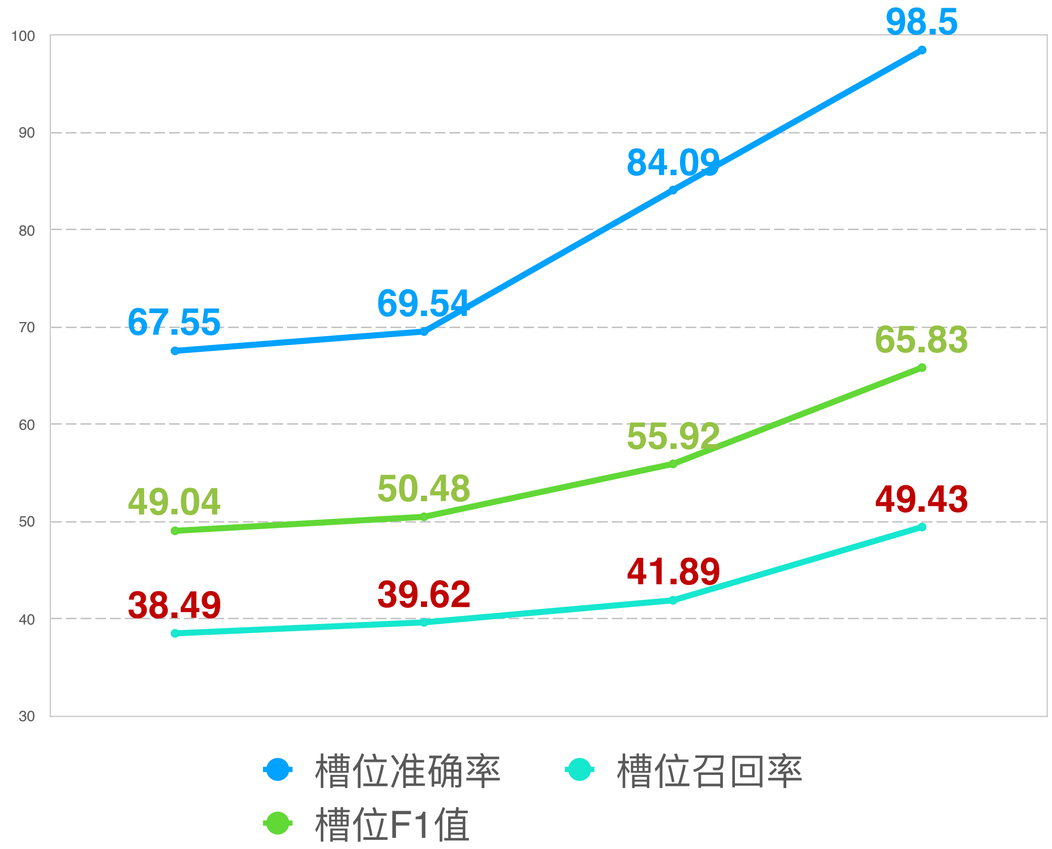
NLU
质量指标


F1值
F-measure
F1= (2*准确率 * 召回率)
/ (准确率+ 召回率)
NLU
召回率
Recall
识别为正确的样本数/
所有样本中正确的数
准确率
Precision
识别为正确的样本数/
识别出的数

AI

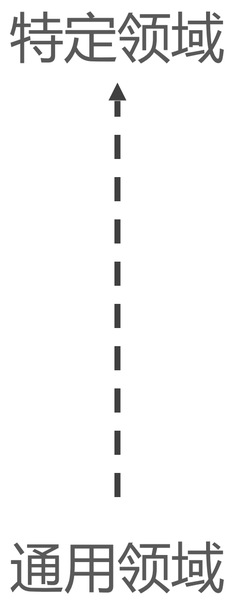
NLU效果评估


100
87.5
75
62.5
50
92.77
86.39
75.71
74.01
73.16
68.87
62.04
60.65
60.39
57.25
52.55
51.37
意图准确率
意图召回率
意图F1值


100
90
80
70
60
50
40
30
98.5
84.09
69.54
67.55
65.83
55.92
50.48
49.04
49.43
41.89
39.62
38.49
槽位准确率
槽位召回率
槽位F1值

AI

对话系统



任务型对话

- 机器人
- 智能音箱

QA问答

- 手机语音助手
- 车载语音系统

开放域聊天

- 小冰
- Siri
- 机器人
- 智能音箱

AI

对话系统

用户Query+Context

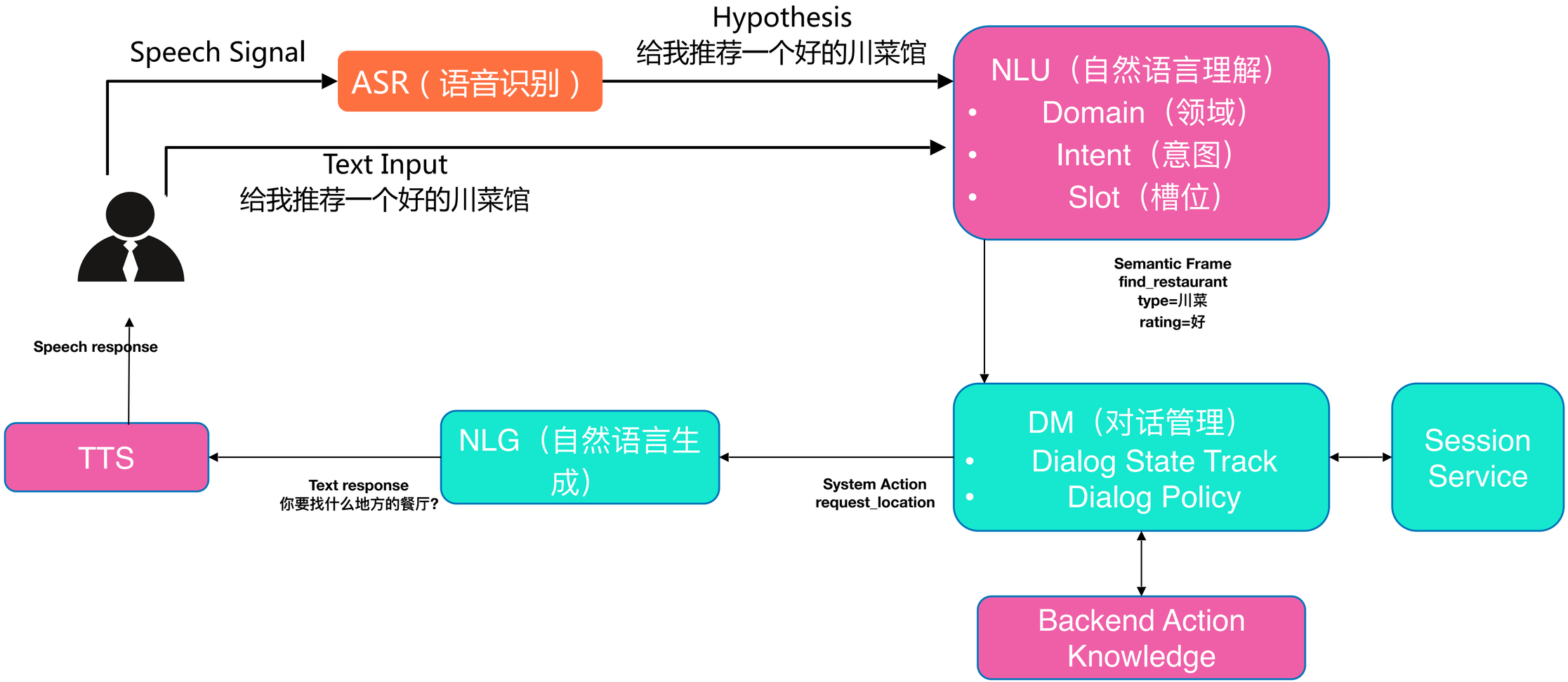
意图识别

中控系统

Chat-Bot　QA-Bot　Task-Bot　Other-Bots

AI

任务型对话系统

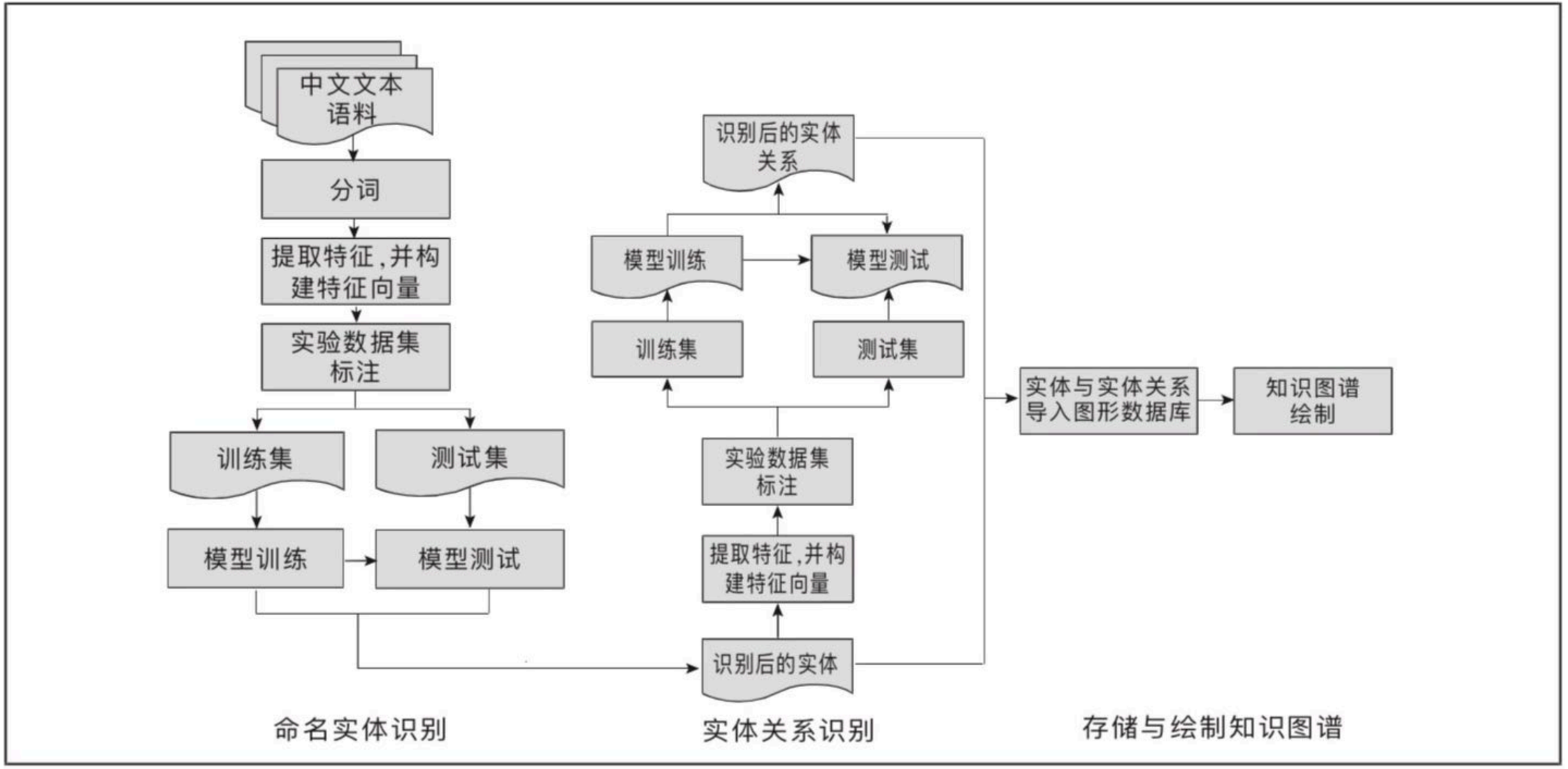

Speech Signal
ASR（语音识别）
Hypothesis
给我推荐一个好的川菜馆
NLU（自然语言理解）
• Domain（领域）
• Intent（意图）
• Slot（槽位）
Text Input
给我推荐一个好的川菜馆
Semantic Frame
find_restaurant
type=川菜
rating=好
Speech response
TTS
NLG（自然语言生成）
Text response
你要找什么地方的餐厅?
System Action
request_location
DM（对话管理）
• Dialog State Track
• Dialog Policy
Session Service
Backend Action Knowledge

DM

质量指标

闲聊型	任务型	问答型
① CPS（Conversations Per Session，平均单次对话轮数）	① 留存率：可发现用户有没有形成使用习惯	①最终求助人工比例
② 相关性和新颖性	② 完成率：用户任务达成率	②重复问同样问题的比例
③ 话题终结者	③ 相关指标：每个任务评价slot填入轮数或填充完整度	③“没答案”之类的比例

基于深度学习算法的商业知识图谱构建流程



KG
质量指标

KG

knowledge Graph

准确率

覆盖率

AI应用孵化器

效果指标-知识图谱效果

北京	属性名	属性值	覆盖率
商品总数：51500 抽取总数：51500 upc缺失：5116	品牌	蒙牛等82个	96.12%(49504)
	规格	ml	86.94%(44775)
	包装	盒，瓶，袋，罐，包	54.67%(28157)
	口味（多）	草莓等19个	26.39%(13592)
	营养成分	全脂，低脂，脱脂	10.59%(5456)

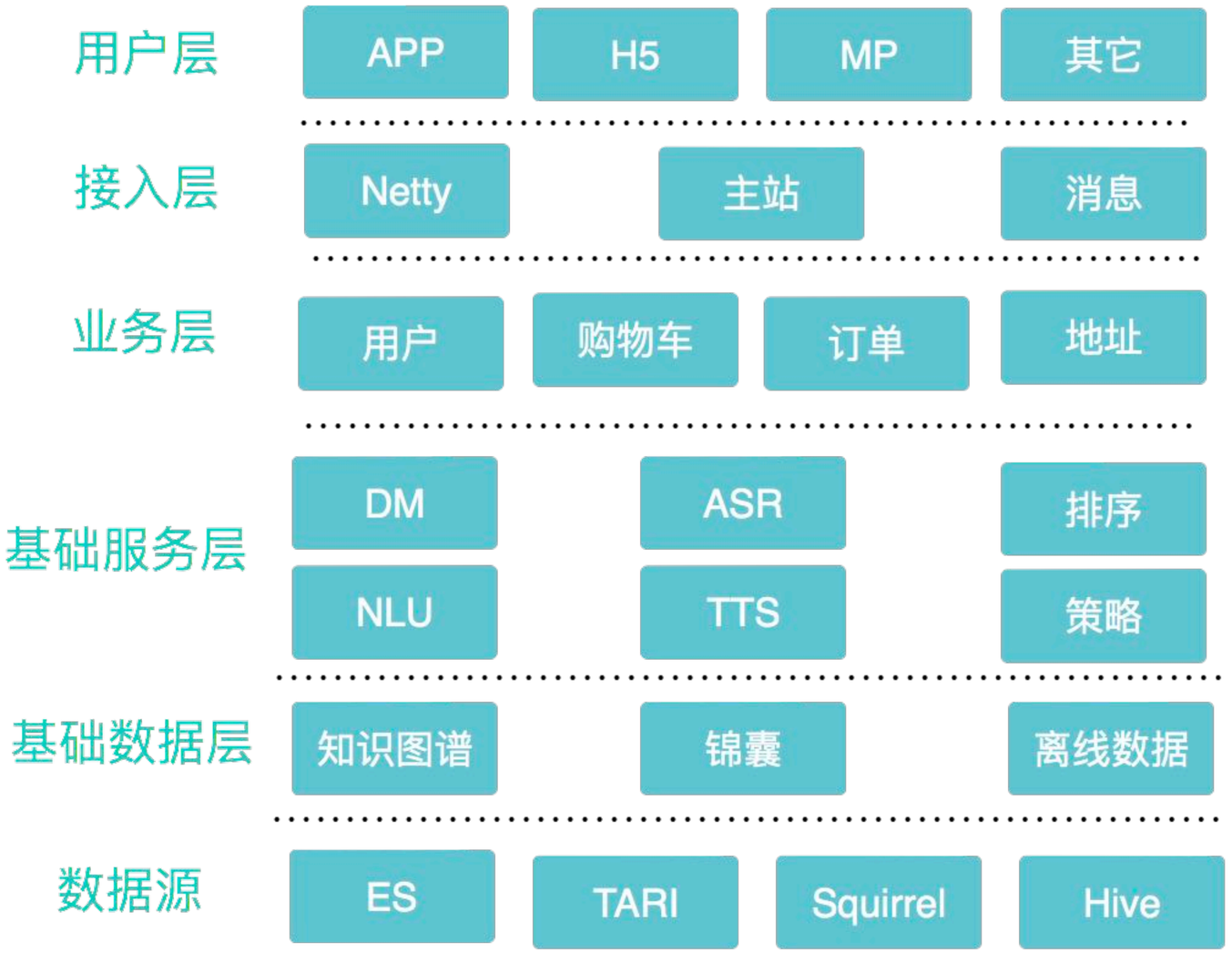
全量	属性名	属性值	覆盖率
商品总数：2246104 抽取总数：2246104 upc缺失：515597	品牌	蒙牛等82个	87.28%(1960508)
	规格	ml	78.98%(1773886)
	包装	盒，瓶，袋，罐，包	69.02%(1550247)
	口味（多）	草莓等19个	31.44%(706192)
	营养成分	全脂，低脂，脱脂	4.55%(102207)

	品类	品牌	规格	单重	数量	包装	口味	营养成分
0	0	0	31	31	31	64	137	185
1	198	181	160	161	160	135	56	15
2	2	0	1	0	1	1	0	0
3	0	19	8	8	8	0	7	0
准确率	99.00%	90.50%	95.50%	96.00%	95.50%	99.50%	96.50%	100.00%

商超领域，知识图谱完成初次数据覆盖度、数据准确率评估，覆盖28个品类，10个属性，图例为牛奶品类，数据覆盖度、数据准确率情况

AI

语音产品核心



01 ASR

将声音转化为文字的过程,相当于人类的耳朵

02 NLU

让计算机能够理解和生成人类语言,包括分词、词性分析、信息抽取等，相当于人类的大脑

03 DM

负责聊天管理和体验，即维护和更新对话的状态，以及对话状态，做出决策，相当于人类的大脑

04 TTS

将文字转化为声音，相当于人类的嘴巴

AI

效果指标-实现策略

01 按优先级

按重要程度

优先实现核心业务的效果指标

02 按实现成本

按实现难度、时间成本

优先实现成本低收益高的效果指标

03 采集数据策略

新项目PM 提供输入数据，QA补充

随着产品迭代，扩充真实用户输入

指标	评价公式	优先级	成本
平均对话轮数	完成任务的平均交互次数	★★	★★
任务完成率	达成用户意图数/用户意图数	★★	★★
没答案比例	没答案数/样本数	★★	★★
唤醒率	呼唤次数/被成功唤醒次数	★	★
识别率	（插入词数+替换词数+删除词数）/标准词	★	★★
召回率	正确数/样本数	★★★	★
准确率	正确数/识别数	★★★	★
F1值	(2*准确率 * 召回率) / (准确率+ 召回率)	★★★	★
用户任务达成率	同任务完成率	★★	★★
对话交互效率	跳出DM现有交互流程次数/总交互次数	★★	★★
数据准确率	正确数/识别数	★★	★★
数据覆盖度	分维度已有数据属性/全网分维度数据属性	★★★	★★

Q&A

二分类问题常用指标：
https://blog.csdn.net/quincuntial/article/details/69596456
https://blog.csdn.net/zhihua_oba/article/details/78677469

书单

- 科学的极致：漫谈人工智能
- 走进2050，注意力、互联网与人工智能
- 与机器人共舞
- 智能时代
- 人工智能：一种现代的方法
- 人工智能 智能系统指南

公众号

- 机器之心
- 全球人工智能
- 哈工大SCIR
- Ai科技评论

CODE A BETTER LIFE

一 行 代 码　　亿 万 生 活

更多技术干货
欢迎关注“美团技术团队”

招聘：测试开发岗位
邮箱：yuhaisheng@meituan.com

美团