**研究报告**

**Siri揭秘——智能问答系统**

**张帅豪**

**2020年12月**

目录

[1 Siri系统 4](#_Toc57623347)

[1.1 背景介绍 4](#_Toc57623348)

[1.2产品定义 4](#_Toc57623349)

[ 服务定义 4](#_Toc57623350)

[ 发展历史 4](#_Toc57623351)

[ 应用设备 5](#_Toc57623352)

[ 支持语言 5](#_Toc57623353)

[1.3现状分析 5](#_Toc57623354)

[ Siri准确度 5](#_Toc57623355)

[ Siri市场份额 6](#_Toc57623356)

[1.4相关技术 7](#_Toc57623357)

[ 语音识别技术 7](#_Toc57623358)

[ 前端方面 12](#_Toc57623359)

[ 后台技术 12](#_Toc57623360)

[ 网页搜索 12](#_Toc57623361)

[ 知识计算 12](#_Toc57623362)

[1.5对比分析 13](#_Toc57623363)

[ Siri和华为小e 13](#_Toc57623364)

[ Siri和小爱 14](#_Toc57623365)

[1.6未来展望 16](#_Toc57623366)

[ 1．语音搜索的增加 16](#_Toc57623367)

[ 2．对聊天机器人的需求增加 16](#_Toc57623368)

[ 3．汽车中的语音控制 16](#_Toc57623369)

[ 4．语音助手功能的改进 16](#_Toc57623370)

[ 5．语音控制的可穿戴设备 17](#_Toc57623371)

[ 6．个性化的体验 17](#_Toc57623372)

[ 7．强大的工作场所助力 17](#_Toc57623373)

[2 智能问答系统 18](#_Toc57623374)

[2.1基本定义 18](#_Toc57623375)

[ 背景介绍 19](#_Toc57623376)

[ 服务定义 20](#_Toc57623377)

[ 服务场景 21](#_Toc57623378)

[ 发展历史 22](#_Toc57623379)

[ 基本分类 23](#_Toc57623380)

[ 现有成果 25](#_Toc57623381)

[2.3基础理论 26](#_Toc57623382)

[ 语音识别ASR 27](#_Toc57623383)

[ 回答状态跟踪 28](#_Toc57623384)

[ 自然语言生成NLG 36](#_Toc57623385)

[ 语音合成TTS 38](#_Toc57623386)

[ 问题解析： 42](#_Toc57623387)

[ 答案抽取： 42](#_Toc57623388)

[2.4研究现状 42](#_Toc57623389)

[2.5对比分析 44](#_Toc57623390)

[ 苹果 Siri 45](#_Toc57623391)

[ 微软小冰和小娜 45](#_Toc57623392)

[ IBM Watson 系统 45](#_Toc57623393)

[ Facebook Messenger 45](#_Toc57623394)

[ 亚马逊Alexa 45](#_Toc57623395)

[ Google Assistant 45](#_Toc57623396)

[ 百度度秘 46](#_Toc57623397)

[ 搜狗汪仔 46](#_Toc57623398)

[ 阿里小蜜 46](#_Toc57623399)

[ 科大讯飞灵犀 46](#_Toc57623400)

[ 其它 46](#_Toc57623401)

[2.6未来展望 46](#_Toc57623402)

# Siri系统

## 1.1 背景介绍

Siri的学术化全称为“个人自动智能助理”（Personalized Intelligent Assistant），并非苹果公司自主研发，最初是苹果AppStore的一款应用，苹果公司发现后很快于2010年4月份从斯坦福研究机构国际智能中心（SRI）处收购而来，传言收购价格在2亿美金左右，之后很快Siri从AppStore下架并集成进入IPhone 4S。

## 1.2产品定义

### 服务定义

Siri（Speech Interpretation and Recognition Interface，语音解析及识别接口）是一款内置在苹果iOS系统中的人工智能助理软件。此软件使用自然语言处理技术，用户可以使用自然的对话与手机进行交互，完成搜索资料、查询天气、设置手机日历、设置闹铃等许多服务。“Siri”一词在挪威语中它的意思是“带领你走向胜利的美丽女神”。

### 发展历史

|  |  |
| --- | --- |
| iPhone 4S iOS 5 Siri首次推出 | 功能十分不完善，常被当做玩偶来戏弄 |
| iOS 6 Siri新增15个国家和地区语言 | 根据蒙斯特的测试，在iOS 6中，Siri能够理解91%的问题，而在iOS 5系统中为89%；在iOS 6中，Siri回答问题的准确率77%，高于iOS 5的76%。 |
| iOS7中除了Siri“测试版”字样 | 减少对Google的依赖，优化声音和中文识别，其中在Siri方面去除了“测试版”字样。在此之前一直是处于“测试阶段” |
| iOS8增加唤醒“嘿Siri” |  |
| iOS9中Siri增加了彩色元素 | 在这之前Siri一直是黑白搭配，以智能穿戴为中心的智能家居正在形成 |
| iOS 10 Siri首次开放了权限给第三方应用 | 第一次被移植到苹果的笔记本电脑上，另外向第三方开发商开放，eg：WhatsApp、领英、Pinterest、Square Cash 和 Slack |

### 应用设备

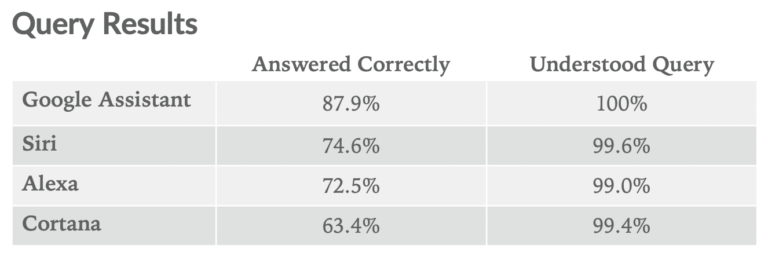
|  |  |
| --- | --- |
| iphone | iPhone 4S iPhone 5 iPhone 5C iPhone 5S iPhone 6  iPhone 6 Plus iPhone 6S iPhone 6S Plus iPhone SE  iPhone 7 iPhone 7 Plus iPhone 8 iPhone 8 Plus  iPhone X iPhone XS iPhone XS Max iPhone XR  iPhone 11 iPhone 11 Pro iPhone 11 Pro Max iPhone SE  iPhone 12 iPhone 12 Mini iPhone 12 Pro iPhone 12 Pro Max |
| ipad | iPad3 iPad4 iPad mini iPad mini 2 iPad mini 3  iPad mini 4 iPad mini 5 iPad Air iPad Air 2  iPad Air 3 iPad Pro |
| ipod | iPod touch 5 iPod touch 6 iPod touch 7 HomePod |
| Apple Watch | 任何版本均支持Siri |
| iphone | iPhone 4S iPhone 5 iPhone 5C iPhone 5S iPhone 6  iPhone 6 Plus iPhone 6S iPhone 6S Plus iPhone SE  iPhone 7 iPhone 7 Plus iPhone 8 iPhone 8 Plus  iPhone X iPhone XS iPhone XS Max iPhone XR  iPhone 11 iPhone 11 Pro iPhone 11 Pro Max iPhone SE  iPhone 12 iPhone 12 Mini iPhone 12 Pro iPhone 12 Pro Max |
| ipad | iPad3 iPad4 iPad mini iPad mini 2 iPad mini 3  iPad mini 4 iPad mini 5 iPad Air iPad Air 2  iPad Air 3 iPad Pro |
| ipod | iPod touch 5 iPod touch 6 iPod touch 7 HomePod |
| Apple Watch | 任何版本均支持Siri |

### 支持语言

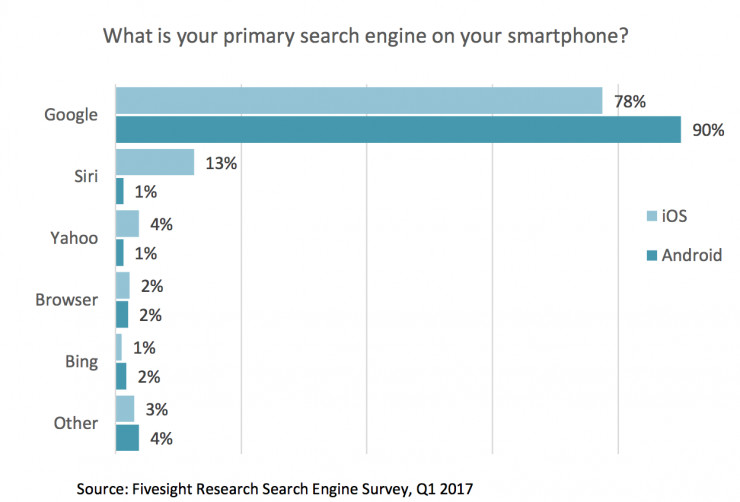
截至IOS6.1，Siri支持中文（普通话 - 中国），中文（粤语 - 中国香港），德文（德国），德文（瑞士），意大利文（意大利），意大利文（瑞士），日文，法文（加拿大），法文（法国），法文（瑞士）英文（澳大利亚），英文（美国），英文（英国），英文（澳大利亚），西班牙文（墨西哥），西班牙文（美国），西班牙文（西班牙）和韩文。

## 1.3现状分析

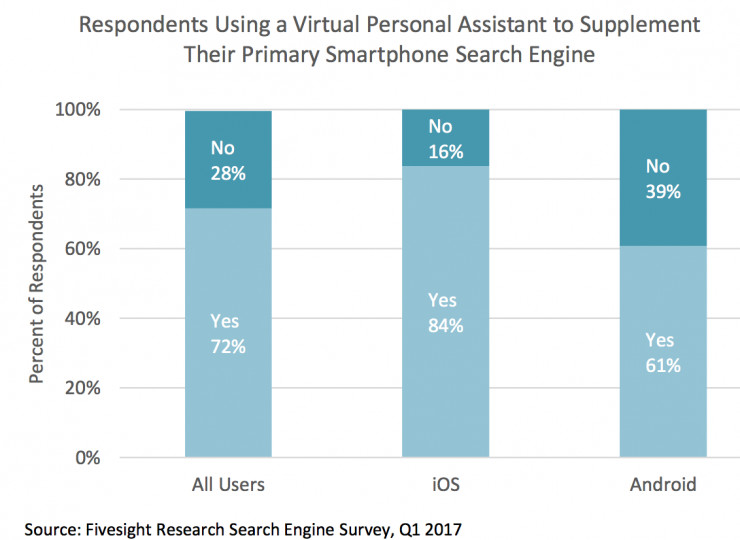
### Siri准确度



### Siri市场份额



在iOS平台上，首先使用Google的人占了78%，而在Android平台上，这个比例高达90%。



72%的人在Android不使用语音助手；在iPhone上，只有16%的人不使用

在这个报告中，我们很明显的可以看到ios携带的siri在市场份额上，明显比除了google外的语音助手更有竞争力。

## 1.4相关技术

Siri首先会根据用户输入的信息在数据库里选择匹配的内容，然后经过一定算法处理以后返回给用户。这里主要涉及以下技术：语音识别技术，负责将语音转化为文本；数据库，负责组织一定形式的数据，便于机器人提取知识；对于Siri来说，它的数据库并不在本地，而是分布式的云端，通过云计算技术（以google为代表的网页搜索技术；②以 Wolfram Alpha 为代表的知识搜索技术；③以Wikipedia 为代表的知识库技术（包括其他百科，如电影百科等）；④以Yelp 为代表的问答以及推荐技术。），实现数据的整合和提取。和Alice等机器人不同的是，Siri似乎可以抛弃繁琐的语法结构，甚至思维模式也可以混乱，Siri会结合上下文结构去理解，它还会从人类语言史的角度出发，利用人工智能系统去分析，并在绝大多数情况下领会你的意思。 在这一点上，Siri显得更为智能，因为它更像人工智能而不是聊天机器人。

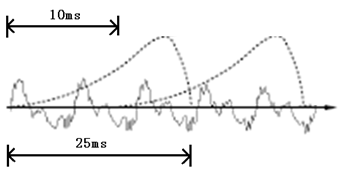
### 语音识别技术

首先，我们知道声音实际上是一种波。常见的mp3等格式都是压缩格式，必须转成非压缩的纯波形文件来处理，比如Windows PCM文件，也就是俗称的wav文件。wav文件里存储的除了一个文件头以外，就是声音波形的一个个点了。下图是一个波形的示例。



在开始语音识别之前，有时需要把首尾端的静音切除，降低对后续步骤造成的干扰。这个静音切除的操作一般称为VAD，需要用到信号处理的一些技术。

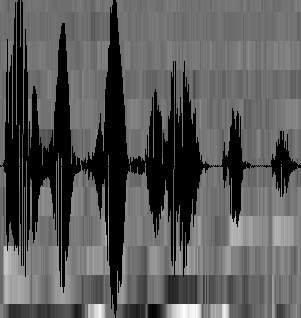
要对声音进行分析，需要对声音分帧，也就是把声音切开成一小段一小段，每小段称为一帧。分帧操作一般不是简单的切开，而是使用移动窗函数来实现，这里不详述。帧与帧之间一般是有交叠的，就像下图这样：



图中，每帧的长度为25毫秒，每两帧之间有25-10=15毫秒的交叠。我们称为以帧长25ms、帧移10ms分帧。

分帧后，语音就变成了很多小段。但波形在时域上几乎没有描述能力，因此必须将波形作变换。常见的一种变换方法是提取MFCC特征，根据人耳的生理特性，把每一帧波形变成一个多维向量，可以简单地理解为这个向量包含了这帧语音的内容信息。这个过程叫做声学特征提取。实际应用中，这一步有很多细节，声学特征也不止有MFCC这一种，具体这里不讲。

至此，声音就成了一个12行（假设声学特征是12维）、N列的一个矩阵，称之为观察序列，这里N为总帧数。观察序列如下图所示，图中，每一帧都用一个12维的向量表示，色块的颜色深浅表示向量值的大小。

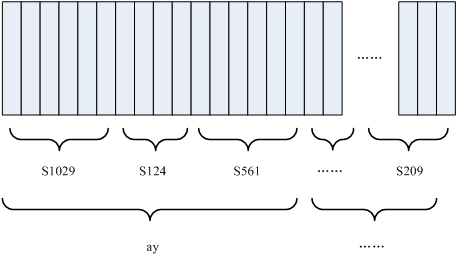


接下来就要介绍怎样把这个矩阵变成文本了。首先要介绍两个概念：

1. 音素：单词的发音由音素构成。对英语，一种常用的音素集是卡内基梅隆大学的一套由39个音素构成的音素集，参见[The CMU Pronouncing Dictionary](https://link.zhihu.com/?target=http%3A//www.speech.cs.cmu.edu/cgi-bin/cmudict" \t "_blank)‎。汉语一般直接用全部声母和韵母作为音素集，另外汉语识别还分有调无调，不详述。
2. 状态：这里理解成比音素更细致的语音单位就行啦。通常把一个音素划分成3个状态。

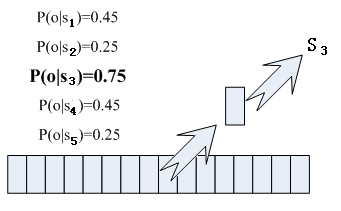
语音识别是怎么工作的呢？实际上一点都不神秘，无非是：  
把帧识别成状态（难点）。  
把状态组合成音素。  
把音素组合成单词。

如下图所示：



图中，每个小竖条代表一帧，若干帧语音对应一个状态，每三个状态组合成一个音素，若干个音素组合成一个单词。也就是说，只要知道每帧语音对应哪个状态了，语音识别的结果也就出来了。

那每帧音素对应哪个状态呢？有个容易想到的办法，看某帧对应哪个状态的概率最大，那这帧就属于哪个状态。比如下面的示意图，这帧在状态S3上的条件概率最大，因此就猜这帧属于状态S3。



那这些用到的概率从哪里读取呢？有个叫“声学模型”的东西，里面存了一大堆参数，通过这些参数，就可以知道帧和状态对应的概率。获取这一大堆参数的方法叫做“训练”，需要使用巨大数量的语音数据，训练的方法比较繁琐，这里不讲。

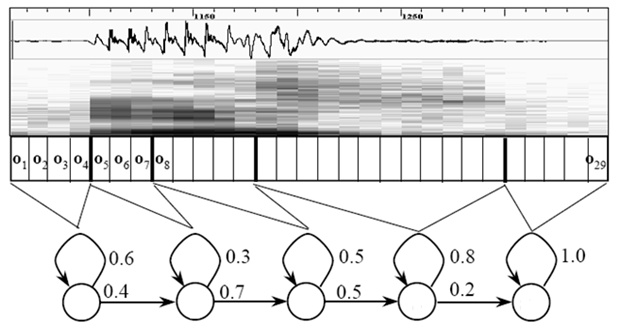
但这样做有一个问题：每一帧都会得到一个状态号，最后整个语音就会得到一堆乱七八糟的状态号。假设语音有1000帧，每帧对应1个状态，每3个状态组合成一个音素，那么大概会组合成300个音素，但这段语音其实根本没有这么多音素。如果真这么做，得到的状态号可能根本无法组合成音素。实际上，相邻帧的状态应该大多数都是相同的才合理，因为每帧很短。

解决这个问题的常用方法就是使用隐马尔可夫模型（Hidden Markov Model，HMM）。这东西听起来好像很高深的样子，实际上用起来很简单：  
第一步，构建一个状态网络。  
第二步，从状态网络中寻找与声音最匹配的路径。

这样就把结果限制在预先设定的网络中，避免了刚才说到的问题，当然也带来一个局限，比如你设定的网络里只包含了“今天晴天”和“今天下雨”两个句子的状态路径，那么不管说些什么，识别出的结果必然是这两个句子中的一句。

那如果想识别任意文本呢？把这个网络搭得足够大，包含任意文本的路径就可以了。但这个网络越大，想要达到比较好的识别准确率就越难。所以要根据实际任务的需求，合理选择网络大小和结构。

搭建状态网络，是由单词级网络展开成音素网络，再展开成状态网络。语音识别过程其实就是在状态网络中搜索一条最佳路径，语音对应这条路径的概率最大，这称之为“解码”。路径搜索的算法是一种动态规划剪枝的算法，称之为Viterbi算法，用于寻找全局最优路径。



这里所说的累积概率，由三部分构成，分别是：

1. 观察概率：每帧和每个状态对应的概率
2. 转移概率：每个状态转移到自身或转移到下个状态的概率
3. 语言概率：根据语言统计规律得到的概率

其中，前两种概率从声学模型中获取，最后一种概率从语言模型中获取。语言模型是使用大量的文本训练出来的，可以利用某门语言本身的统计规律来帮助提升识别正确率。语言模型很重要，如果不使用语言模型，当状态网络较大时，识别出的结果基本是一团乱麻。

这样基本上语音识别过程就完成了。

### 前端方面

在前端方面，即面向用户，和用户交互（User Interface）的技术，主要是语音识别以及语音合成技术。语音识别技术是把用户的口语转化成文字，其中需要强大的语音知识库，因此需要用到所谓的“云计算”技术。而语音合成则是把返回的文字结果转化成语音输出，这个技术理论上本地就能完成，但不知道Siri是否如此，当然，在云端完成也并无不可，在当前无线带宽下，那点语音流量根本不算什么。

### 后台技术

这些技术的目的就是处理用户的请求，并返回最匹配的结果，这些请求类型很多，千奇百怪，要处理好并不简单。基本的结构猜测可能是分析用户的输入（已经通过语音转化），根据输入类型，分别采用合适的技术（合适的技术后台）进行处理。这些合适的后台技术包括，①以Google 为代表的网页搜索技术；②以Wolfram Alpha为代表的知识搜索技术（或者知识计算技术）；③以[Wikipedia](https://baike.baidu.com/item/Wikipedia) 为代表的知识库（和 Wolfram Alpha 不同的是，这些知识来自人类的手工编辑）技术（包括其他百科，如电影百科等）；④以[Yelp](https://baike.baidu.com/item/Yelp) 为代表的问答以及推荐技术。 [4]

### 网页搜索

大家天天用的[Google](https://baike.baidu.com/item/Google)就是这个技术的代表，而国内则以百度为代表。

### 知识计算

**（搜索）技术（Computational Knowledge）**

这个技术的代表是WolframAlpha。

不同于搜索互联网信息，Wolfram|Alpha将从公众的(包括公开的网页等)和获得授权的资源中，发掘、建立起一个异常庞大的经过组织的数据库，再利用高级的自然语言算法进行处理，最终构造出一个类似于谷歌搜索的工具。

和网页搜索技术不同的是，在这个系统中，得到的答案结构化程度很高，比如搜索China，能得到和中国相关的各种参数以及资料，并以接近表格的方式呈现。Wolfram|Alpha也能理解部分自然语言，比如输出How old are you，其会回答Wolfram|Alpha的年龄。想测试这项技术的请移步Wolfram|Alpha。

这个技术国内做的应该有，但还没有产品，也许百度的[框计算](https://baike.baidu.com/item/%E6%A1%86%E8%AE%A1%E7%AE%97)算是半个。所以机会大大的有。

**知识库技术**

相比于网页搜索技术，基本以一个词条或者主题为单位，因此得到的数据价值高，知识量大，并且结构化程度好。相比于知识计算技术，这些技术需要人的参与，这有利也有弊，利就是，毕竟暂时人比机器聪明，编辑出来的知识更丰富，准确；弊就是，人力有限，即使像维基那样，发动社区的力量，也不能产生足够的知识，而知识计算，理论上，只需要算法够，是可以产生“无限”的知识的。

**问答推荐技术**

其实这不能称为一个技术，应该属于知识库的技术。不同的是，这个技术针对的是一些生活信息，这些信息的地域化程度很高，典型代表为[Yelp](https://baike.baidu.com/item/Yelp" \t "_blank)。由于这东西比较简单，就不仔细介绍了。

**语言模型技术**

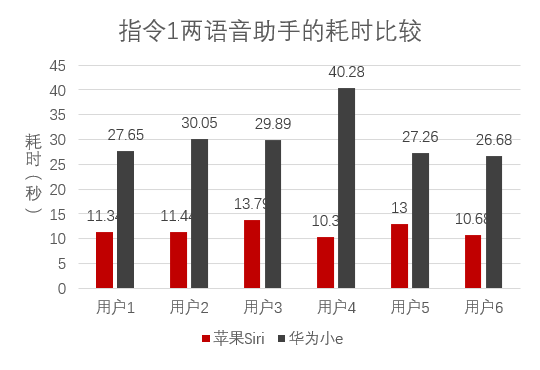
这是如何提升Siri识别企业和餐厅等本地兴趣点名称的方法，融合了用户所在地的知识，也就是所谓的Geo-LM，借此提升Siri的语音识别能力。这些模型让Siri可以更好地评估用户的发音内容。

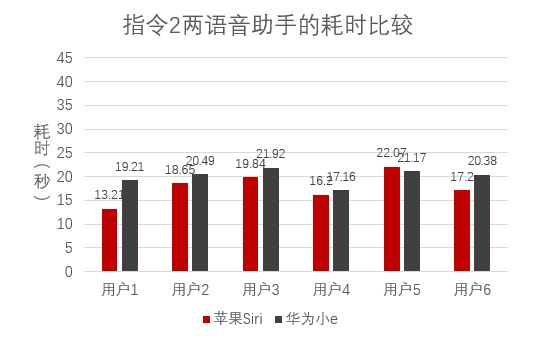
## 1.5对比分析

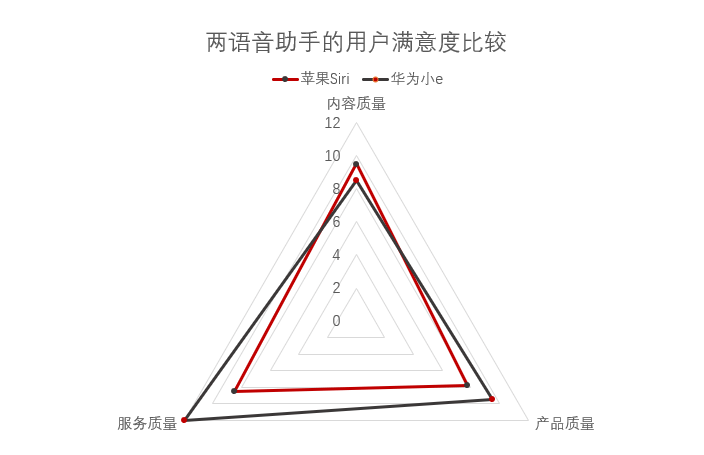
### Siri和华为小e

第1组：“给妈妈发微信祝她生日快乐”

第2组：“发短信给妈妈说记得提醒我明天下午两点去小剧场彩排节目”





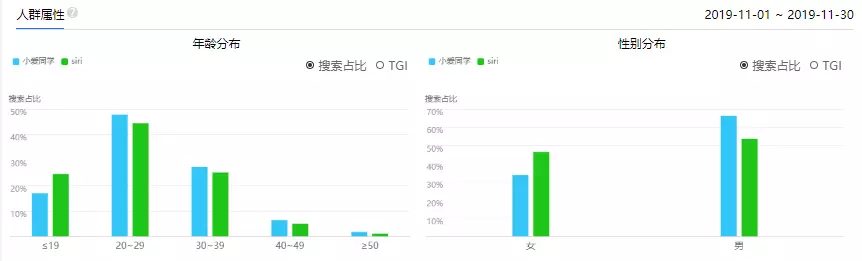


### Siri和小爱

**产品战略对比**

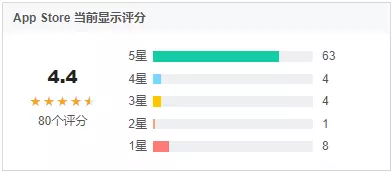


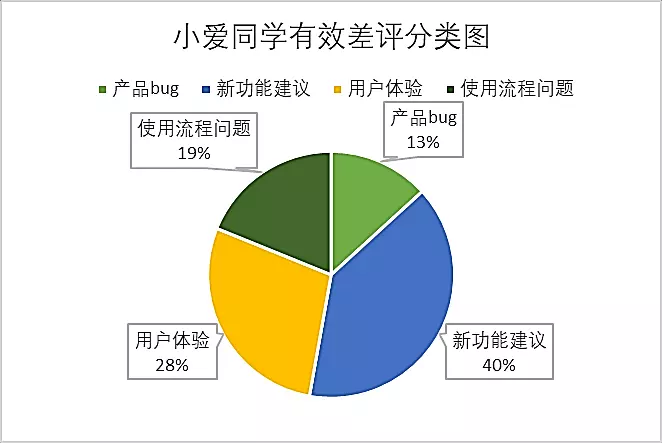
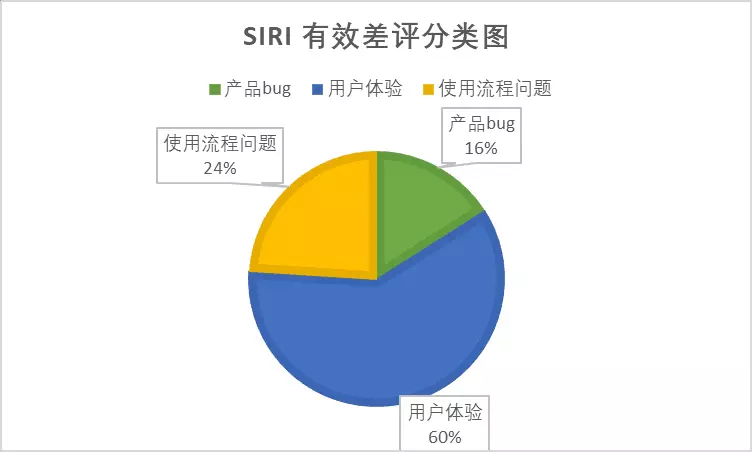
**用户数据**



**用户差评反馈**







## 1.6未来展望

### 1．语音搜索的增加

随着语音助手在回答人们的问题上越来越出色，越来越多的人会更喜欢使用它们。口头询问的吸引力显而易见。口头提出问题通常比输入问题容易。使用语音命令更快，并且允许用户在双手忙时进行搜索。在多任务处理时代，免提技术越多越好。

### 2．对聊天机器人的需求增加

聊天机器人是一种越来越普遍的网站功能。如果数字助理可以用自然语言回答访客的口头请求，那么查询结果将大有帮助。支持AI的聊天机器人将把客户服务提升到一个全新的水平。

### 3．汽车中的语音控制

到2020年，AI助手将进入越来越多的产品领域。语音控制系统最有前途的应用之一是在汽车上。不仅口头功能在新车中更为常见，而且很快就会变得越来越重要。

### 4．语音助手功能的改进

智能音箱已经有了一系列令人印象深刻的功能。他们可以关掉你的灯，给你的联系人发短信，用笑话和歌曲来娱乐你。这些设备的日益普及将吸引更多的公司与它们合作，改进它们本已引人注目的功能。

这种合作对各方都是双赢的局面。企业可以通过智能音箱进入人们的家庭中。智能音箱厂家可以看到销量的增加，让客户得到更方便、更有帮助的体验。

### 5．语音控制的可穿戴设备

虚拟助理也开始走出家门。随着人们习惯了智能音箱的便利性，他们会想要随时随地使用这些服务，而制造商将会奋起满足这一需求。语音AI助手将与可穿戴设备融合。

### 6．个性化的体验

随着NLP技术的发展，语音助手体验将变得更加个性化。可以区分多个独特的声音，而这个数字很快将变得微不足道。更加精致的语音识别软件将能够识别更多的音调，并理解不同来源和强度的口音。

个性化程度的提高有几个重大影响。同一座大楼中的多个人可以向数字助理询问他们的日程安排或提醒，并获得相关的准确答案，而无需确认自己。智能设备将学习不同用户的偏好，并以个性化的方式提供个性化信息。

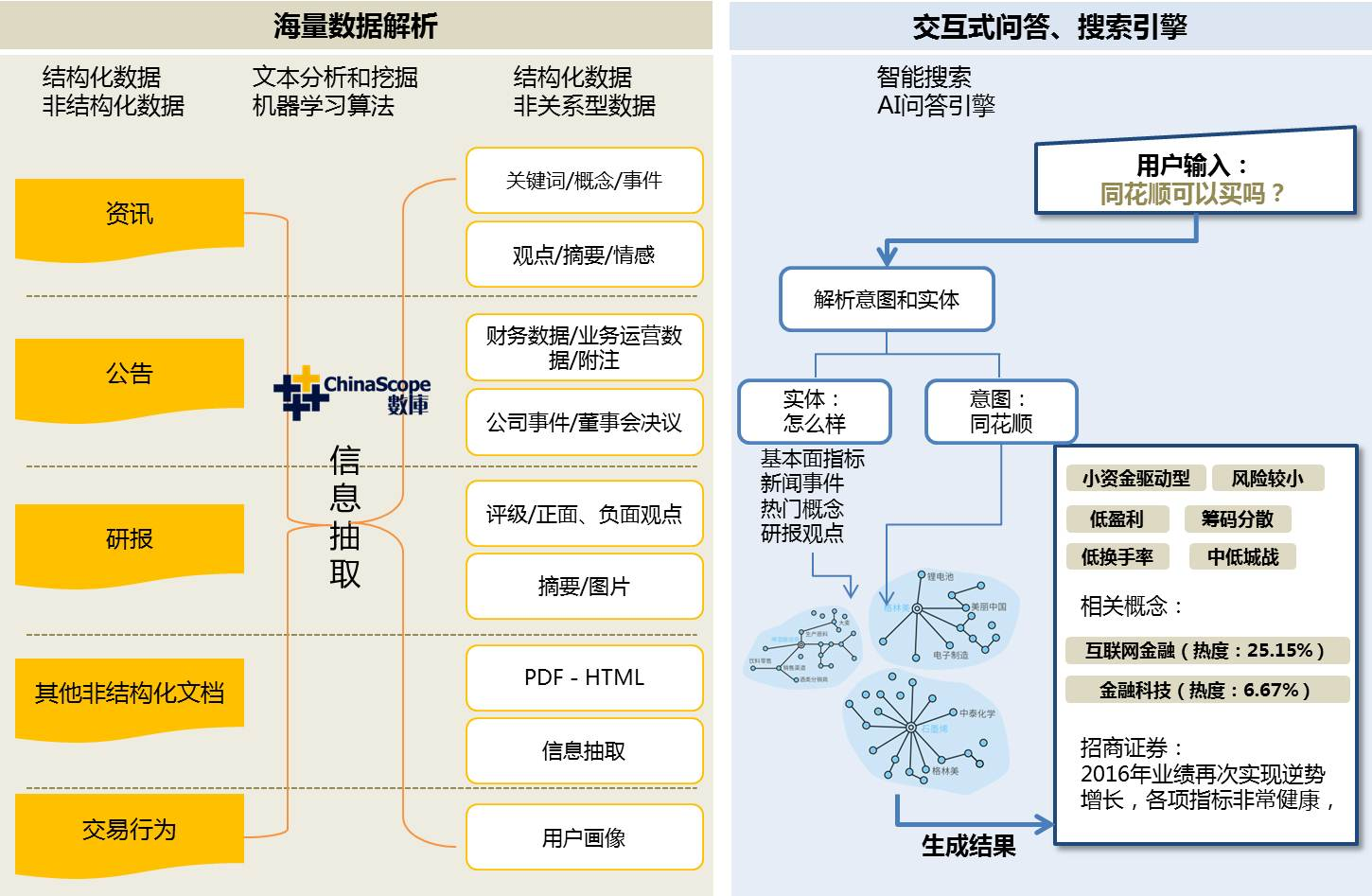
### 7．强大的工作场所助力

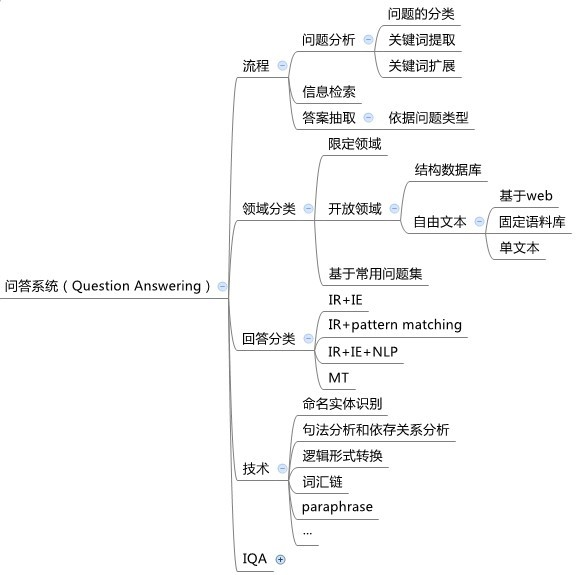
在未来的一年里，AI助手将从家庭使用转向办公室使用。语音技术可能会给商业智能带来一场革命，而且它已经开始这样做了。微软已经将其助手Cortana的重心转移，将其改进为一种办公工具。

经过优化的专业助理可以管理电子邮件服务，处理整个工作场所的日程安排并为员工提供重要的业务数据。

# 智能问答系统

## 2.1基本定义





### 背景介绍

随着人工智能技术的发展，智能服务系统框架内的智能问题系统也取得了诸多应用及发展，尤其是在应用场景、技术标准、发展框架下，以用户信息及知识需求解决为导向，解决用户在服务层面的信息及知识需求。在智能问答系统的应用框架内，其应用知识领域主要有开放、限定两种模式，开放模式主要是按照开放的问答系统应用框架模式，对客户的提问格式、方式均不限定，以关联性的模式获取答案的反馈。

以某垂直领域电商企业客服中心为例，该中心主要是自营垂直电商业务为主，电商服务与人们的生活息息相关，电商资讯的服务需求较多，用户在咨询中可以先通过开放式的智能问答模式，在降低人工客服的压力同时提升服务的效率。在技术应用语言层面，开放式的问答技术模式基本都是按照自然语言的方式，按照限定知识模式进行技术应用。在限定自然语言的处理过程中，自然语言处理任务在垂直电商领域应用还需要按照标准训练模式、数据构建模式来设计智能问答系统的应用路径。

### 服务定义

问答系统是信息检索系统的一种高级形式，它可以通过链接知识库方式检索到用户问题的答案。问答系统与信息检索中的语义搜索有点类似，把用户输入的问题转换为一个有结构的语义表达式，然后从知识库中寻找答案，并直接反馈给用户。

问答系统的答案可以从结构化知识库获取，用于规章制度、规则条款等问答形式，也可以通过结合一些已有的知识库和数据库资源，利用百度百科、维基百科等非结构化文本信息，用信息抽取的方法提取关键信息，并构建知识图谱作为问答系统的后台支撑，再结合知识推理等方法为用户提供更深层次语义理解的答案。

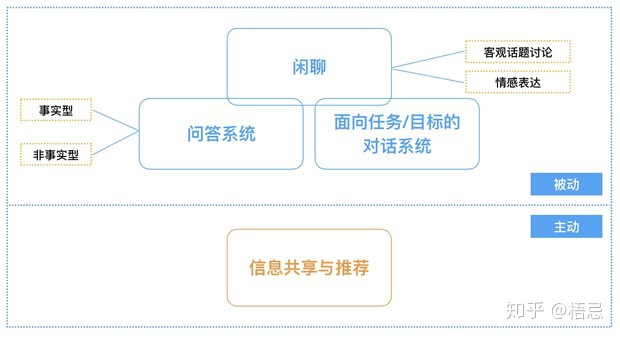
在这里主要讲解的是针对特定领域的知识进行一问一答，也称之为单轮对话、问答对或FAQ，即直接根据用户的问题给出精准的答案。

Eg客服智能问答系统涉及领域非常广阔，其主要从用户意图、期望动作、应用策略进行设计，按照不同问答数据来源进行区分，客服智能问答系统主要可以分为如下三种类型。一是基于自由文本的客服智能问答系统，该系统主要是从问句主题、分类两个部分来进行拆分及分析，该系统通过匹配模式、机器学习、分类应用、主题识别、句法分析的模式确认。后续，按照文档检索的方式，选择合适的答案反馈给客户。

二是基于问答的客服智能问答系统，该系统主要按照常见list列表（FAQ）的智能模式以及社区辅助问答（CQA）和常用问题列表（FAQ）的智能问答模式，两个系统之间的相关性主要是按照问句、模板数据模式计算相关性，将评分相对较高的预设模板通过问答系统提供给客户。

三是基于结构化数据问答模式建立的问答系统，该系统主要是通过分析的模式，先把问句转化为查询，简单的查询模式会在结构化的数据库中自动形成检索答案，不过答案缺少问句转化以及深层次的语义分析，与精准答案的匹配度较低。另外一种方法是按照深度学习模式下构建的问答系统模式，先经过分词的方式，定向输入词语，通过模型加以识别、链接、关系等，最终获得匹配度较高的答案。

### 服务场景



问答系统(Question Answering System，QA System)是用来回答人提出的自然语言问题的系统。从狭义上来说，问答系统是聊天机器人的其中一个组成部分

按照知识领域，可以将问答系统分类为“具体领域”以及“开放领域”。具体领域系统专注于回答特定领域的问题，如医药、体育、政府事务等。开放领域系统则希望不设限问题的内容范围，天文地理无所不问。

按照问题类型，又可作如下划分:

事实型问题：WH 问题，例如 when / who / where 等；

是非型问题：Is Beijing the capital of China?

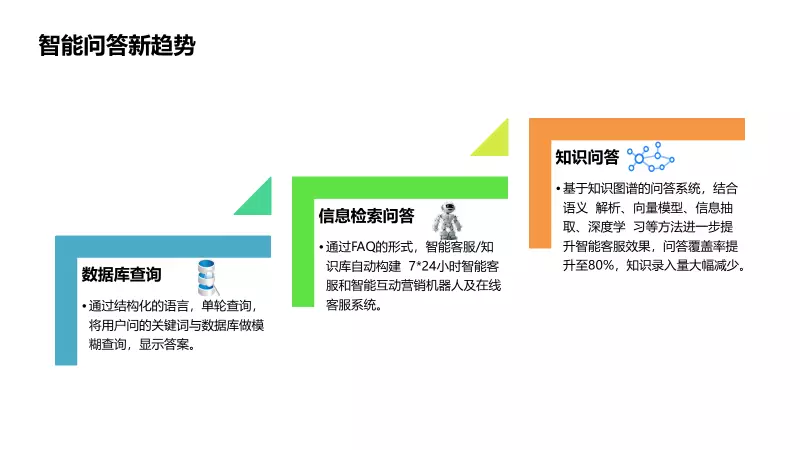
对比型问题：Which city is larger, Shanghai or Beijing?

观点型问题：What is Chinese opinion about Donald Trump?

原因/结果型问题：how / why / what 等。

是非型和对比型这类客观问题以 事实型问题 为核心（对于是非型问题 Is Beijing the capital of China? ，事实型问题 What is the capital of China? 是其回答的基础。对于对比型问题 Which city is larger, Shanghai or Beijing? ，事实型问题 How large is Shanghai? 和 How large is Beijing? 是其回答的基础。），因此我们又可以简单地将问题类型分为 事实型 和 非事实型。针对两类问题类型的解决方案大相径庭。

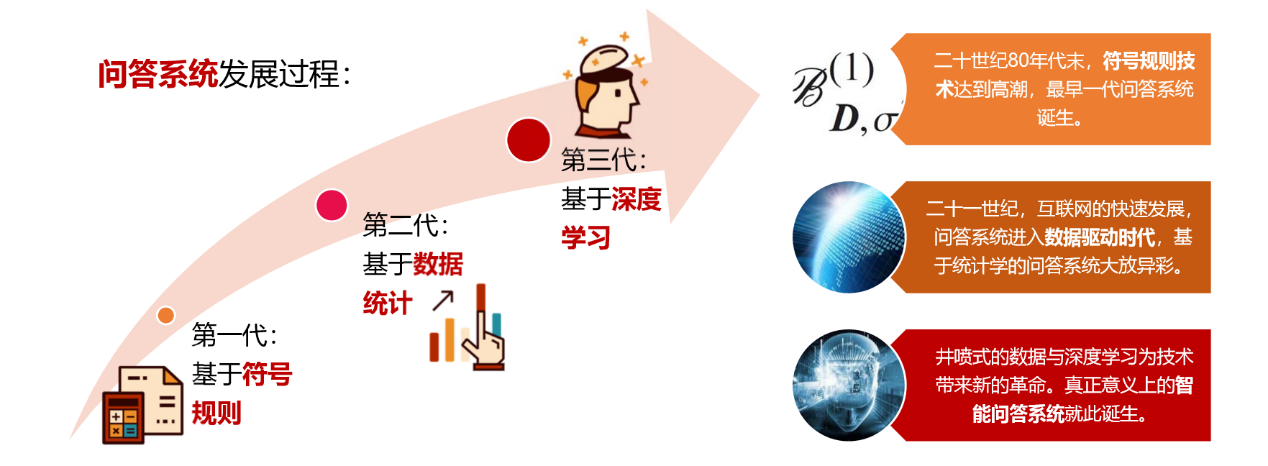
### 发展历史



智能问答系统已经有 70 年的发展历史。早期的智能问答系统大多针对特定的领域而设

计，并且数据量也很有限，不容易进行扩展，例如：Baseball和 Lunar，这些诞生在上世纪五六十年代的智能问答系统通常只接受特定形式的自然语言问句，而且可以供智能问答系统进行训练的数据也很少，所以无法进行基于大数据的开放领域的问答从而未被广泛用。进入九十年代之后，由于互联网的发展，大量可供训练的问答对在网上可以被搜集和找到。尤其是 TREC-QA评测的推出，极大推动促进了智能问答系统的发展，研究人员在该语料库上训练和测试各种问答模型，先后提出了基于逻辑推理的方法，基于模板匹配的方法，基于机器学习的方法和基于数据冗余性的方法[7]等许多领先方法。在此阶段，人们主要利用信息检索或浅层语义理解技术去从大量候选集中寻找答案从而构建智能问答系统，故检索式问答技术取得了巨大发展。但是检索式问答技术存在一个缺陷，就是答案中一定至少包含一个用户问句中含有的字或者词，但是这在实际情况中往往是不成立的。虽然浅层语义理解技术部分解决了这个问题，但是由于用户问句是自然语言，自然语言有着天然的复杂性，由于存在着以上缺陷，检索式问答技术不能真正很好地解决用户的需求。一直以来，阻碍智能问答系统向前发展的两个最主要因素是缺乏高质量的数据和强大的自然语言处理技术，不过随着维基百科，百度百科，搜狗百科这些基于用户协同生成内容的互联网应用的兴起，越来越多的高质量数据被积累和得到。基于此，大量的被精心设计以自动或半自动方式生成的知识库（例如 Freebase、YAGO、DBpedia 等）被建立起来。至于另一个问题，随着统计机器学习方法的兴起，自然语言处理技术各个子领域都取得了很大的进步，可以说阻碍智能问答系统最大的两个问题正在被科研人员逐步解决。近年来，智能问答系统取得了很大的发展和进步，已经有很多智能问答系统产品问世。例如 IBM 研发的智能问答机器人Watson 在美国智力竞赛节目《Jeopardy!》中战胜人了选手，其所拥有的 DeepQA 系统集成了统计机器学习、信息抽取、知识库集成和知识推理等深层技术。苹果公司的 Siri 系统和微软公司的 cortana 分别在 iPhone 手机中和 Windows10 操作系统中都取得了很好的效果。在国内，众多企业和研究团体也推出了很多以智能问答技术为核心的机器人，例如：微软公司的“小冰”、百度公司的“度秘”和中科汇联公司的“爱客服”，我们可以看到，这些机器人不仅提供情感聊天的闲聊功能，而且还能提供私人秘书和智能客服这样的专业功能。这些智能系统的出现标志着智能问答技术正在走向成熟，预计未来还会有更多功能的机器人问世和解决用户的各种需求。

|  |
| --- |
| 神经语言模型（2001） |
| 多任务学习（2008） |
| 词嵌入（2013） |
| 用于自然语言处理的神经网络（2013） |
| 序列到序列模型（2014） |
| 注意力机制（2015） |
| 基于记忆的神经网络（2015） |
| 预训练的语言模型（2018） |



### 基本分类

按照对话轮数可以将智能对话系统分为：单轮对话与多轮对话。  
按照对话的目的性可以分为：任务导向型（task-oriented)对话系统和非任务导向型(non-task-oriented)对话系统(也称为聊天机器人)。

#### 对话轮数

单轮对话

1.什么是单轮对话

单轮对话与传统的问答系统相类似，是智能对话系统的初级应用。一般表现为**一问一答的形式**，用户提出问题或前发出请求，系统识别用户意图，做出回答或执行特定操作。  
**单轮对话也强调自然语言理解，但一般不涉及上下文、指代、省略或隐藏信息**，相对来说实现难度更低，产品的应用也更加成熟可靠。

2.单轮对话的应用场景

**单轮对话主要应用在目标明确且会话行程短的浅服务类项目中**，如电商平台客服机器人提供的产品介绍、订单信息查询、退换货流程介绍等内容。  
**单轮对话的本质在于取代人工工作中高度重复的标准化客户咨询**，可以理解为一个高效率的自助服务帮助文档或知识库，能够帮助用户快速获取信息、提升咨询效率。

多轮对话

1.什么是多轮对话及其应用场景

**多轮对话一般是任务驱动的多轮对话，用户是带着明确的目的的**，如**订餐**，**订票**，**叫车**，**寻找音乐**、**电影**或**某种商品**等比较复杂的需求来，而这中间有很多限制条件，用户并不能一次将任务所需的关键信息一次性说完，说清楚，因此就要分多轮进行QA问答一方面，用户在对话过程中，可以不断修正和完善自己的需求;另一方方面，当用户在陈述需求不够具体和明确时，机器人可以通过询问，澄清和确认来帮助用户寻找满意的结果，并且在任务的驱动下与用户完成日常的交互，以此不断完善对于用户需求的满足。  
因此，**任务驱动的多轮对话不是一个简单的自然语言理解加信息检索的过程，而是一个决策过程，需要机器在对话过程中不断根据当前的状态决策下一步应该采取的最优动作（如：提供结果，询问特定限制条件，澄清或确认需求，等等）从而最有效的辅助用户完成信息或服务获取的任务**。

2.任务驱动的多轮对话系统

#### 对话的目的性

非任务导向型(non-task-oriented)对话系统

所谓非任务导向型对话系统，就是说这一类的对话是没有明确的目的性的对话。非任务导向的对话系统与人类交互，提供合理的回复和娱乐消遣功能，通常情况下主要集中在开放的领域与人交谈。虽然非任务导向的系统似乎在进行聊天，但是它在许多实际应用程序中都发挥了作用。  
一般来说，对于非任务导向型对话系统，目前用的主要是两种方法：  
① **生成方法**，例如**序列到序列模型（seq2seq）**，在对话过程中产生合适的回复，生成型聊天机器人目前是研究界的一个热点，和检索型聊天机器人不同的是，它可以生成一种全新的回复，因此相对更为灵活，但它也有自身的缺点，比如有时候会出现语法错误，或者生成一些没有意义的回复。  
② **基于检索的方法**，从事先定义好的索引中进行搜索，学习从当前对话中选择回复。检索型方法的**缺点**在于它过于依赖数据质量，如果选用的数据质量欠佳，那就很有可能前功尽弃。

任务导向型（task-oriented)对话系统

面向任务的对话系统是对话系统的一个重要分支。**任务导向型对话系统旨在通过分析对话内容提取用户任务，并且帮助用户完成实际具体的任务**，例如帮助用户找寻商品，预订酒店餐厅等。  
这里介绍任务导向型对话系统的两种实现方法：**管道方法**和**端到端方法**。

1.管道方法  
任务导向型对话系统的典型结构包括四个关键组成部分:  
1）**自然语言理解**（Natural Language Understanding，NLU）：它将用户输入解析为预定义的语义槽。  
如果有一个话语，自然语言理解将其映射为语义槽。插槽是根据不同的场景预先定义的。  
2）**对话状态跟踪**（Dialogue State Tracking，DST）  
概括的说，**对话状态跟踪就是根据多轮的对话来确定用户当前的目标（user goal）到底是什么的过程**，它是确保对话系统健壮性的核心组件。它在对话的每一轮次对用户的目标进行预估，管理每个回合的输入和对话历史，输出当前对话状态。这种典型的状态结构通常称为槽填充或语义框架。传统的方法已经在大多数商业实现中得到了广泛的应用，通常采用手工规则来选择最有可能的输出结果。然而，这些基于规则的系统容易出现频繁的错误，因为最可能的结果并不总是理想的。  
因此，最近深度学习采用的方法是使用一个滑动窗口输出任意数量的可能值的概率分布序列。虽然它在一个领域受过训练，但它可以很容易地转移到新的领域。此处运用较多的模型是，multi-domain RNN dialog state tracking models和Neural Belief Tracker (NBT) 。  
3）**对话策略学习**（Dialogue policy learning）  
根据状态跟踪器的状态表示，策略学习是生成下一个可用的系统操作。无论是监督学习还是强化学习都可以用来优化政策学习。监督学习是针对规则产生的行为进行的，在在线购物场景中，如果对话状态是“推荐”，那么触发“推荐”操作，系统将从产品数据库中检索产品。强化学习方法的引入可以对对话策略进行进一步的训练，以引导系统制定最终的策略。在实际实验中，强化学习方法的效果超过了基于规则和监督的方法。  
4）**自然语言生成**(Natural Language Generation，NLG)。 根据对话策略选择操作进行映射并生成相关的回复。

2.端到端方法

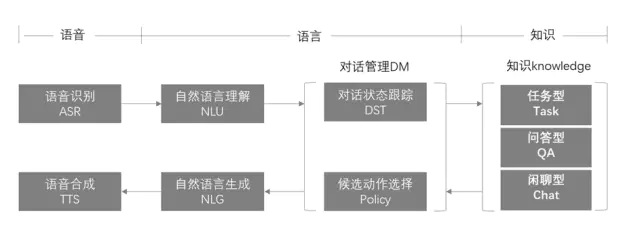
因为基于管道方法的对话系统中有许多特定领域的手工制作，所以它们很难适用于新的领域。近年来，随着端到端神经生成模型的发展，为面向任务的对话系统构建了端到端的可训练框架。与传统的管道模型不同，端到端模型使用一个模块，并与结构化的外部数据库交互。

是一种基于网络的端到端可训练任务导向型对话系统，将对话系统的学习作为学习从对话历史到系统回复的映射问题，并应用encoder-decoder模型来训练。然而，该系统是在监督的方式下进行训练——不仅需要大量的训练数据，而且由于缺乏对训练数据对话控制的进一步探索，它也可能无法找到一个好的策略。  
随着强化学习研究的不断深入，上图的模型首先提出了一种端到端强化学习的方法，在对话管理中联合训练对话状态跟踪和对话策略学习

### 现有成果

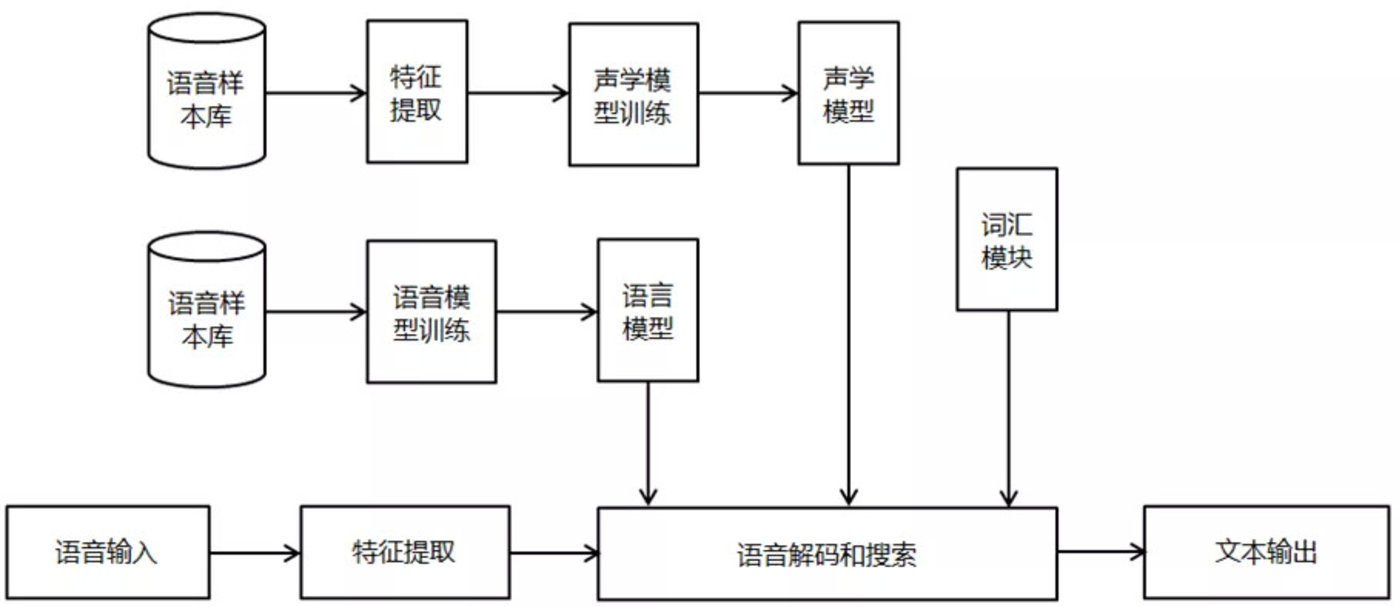
|  |  |
| --- | --- |
| IBM waston | 中国深度学习、大数据分析、云计算、社交商务、物联网、数据管理、企业管理、IT优化、软件开发运维 |
| 科大讯飞 | 1999年成立，语音技术提供商（语音支撑软件、行业应用产品/系统、信息工程和运维服务等），2016年进军智能客服领域 |
| 小i机器人（智臻智能） | 2001年成立，一直都在做智能客服这块，定位智能客服、知识管理、虚拟个人助手、大数据分析、可嵌入式智能芯片软件供应商 |
| 云问（faqRobot） | 2013年成立，智能客服供应商 |
| 捷通华声 | 2000年成立，做语音的，后来转做智能客服，智能图像、智能语音、智能终端、输入法 |
| 智齿科技 | 2013年成立，2C的语音助手起家，后转做智能客服，在线客服系统、2c机器人服务商、工单系统、客服APP |
| 小能机器人 | 2011年成立，在线客服系、智能客服机器人、呼叫中心供应商 |
| 玻森 | 2012年成立，中文语义分析（非交互）提供商 |
| 追一 | 2016年成立，智能客服 |
| 奇智科技 | 2016年成立，app、硬件服务(家教服务及健康养老服务信息系统)供应商 |
| 竹间智能 | 2015年成立，深度学习，智能机器人、情感识别技术供应商 |
| 网易七鱼 | 2015年网易分化的事业部性质，智能客服 |
| 光年无限(图灵) | 2010年成立，智能客服 |
| Udesk | 2013年成立，智能客服 |
| 微软小冰 | 2014年产品发布，聊天机器人 |
| 阿里小蜜 | 2015年产品发布，2C人工智能购物助理虚拟机器人 |
| 百度夜莺 | 2016您产品发布，智能客服平台 |
| 广州极天 | 2007年成立 |
| 小富机器人 | 鼎富科技被神州泰岳收购后发布的产品 |
| 云知声 | 2012年成立，语音转做智能客服 |
| 思必驰 | 2007年成立，智能交互领域的车载服务提供商 |
| 助理来也 | 2015年成立，2C全品类私人助理服务平台 |
| 使爱智能 | 2012年成立，智能客服机器人、语音识别、人脸识别、智能文字识别供应商 |
| 出门问问 | 2014年成立，虚拟个人助理、穿戴式设备供应商 |
| ipsoft | 1998年成立，英文版全球自主服务提供商 |

## 2.3基础理论



### 语音识别ASR

语音识别技术是让机器通过识别把语音信号转变为文本，进而通过理解转变为指令的技术。目的就是给机器赋予人的听觉特性，听懂人说什么，并作出相应的行为。语音识别系统通常由声学识别模型和语言理解模型两部分组成，分别对应语音到音节和音节到字的计算。一个连续语音识别系统（如下图）大致包含了四个主要部分：特征提取、声学模型、语言模型和解码器等。



（1）语音输入的预处理模块

对输入的原始语音信号进行处理，滤除掉其中的不重要信息以及背景噪声，并进行语音信号的端点检测（也就是找出语音信号的始末）、语音分帧（可以近似理解为，一段语音就像是一段视频，由许多帧的有序画面构成，可以将语音信号切割为单个的“画面”进行分析）等处理。

（2）特征提取

在去除语音信号中对于语音识别无用的冗余信息后，保留能够反映语音本质特征的信息进行处理，并用一定的形式表示出来。也就是提取出反映语音信号特征的关键特征参数形成特征矢量序列，以便用于后续处理。

（3）声学模型训练

声学模型可以理解为是对声音的建模，能够把语音输入转换成声学表示的输出，准确的说，是给出语音属于某个声学符号的概率。根据训练语音库的特征参数训练出声学模型参数。在识别时可以将待识别的语音的特征参数与声学模型进行匹配，得到识别结果。目前的主流语音识别系统多采用隐马尔可夫模型HMM进行声学模型建模。

（4）语言模型训练

语言模型是用来计算一个句子出现概率的模型，简单地说，就是计算一个句子在语法上是否正确的概率。因为句子的构造往往是规律的，前面出现的词经常预示了后方可能出现的词语。它主要用于决定哪个词序列的可能性更大，或者在出现了几个词的时候预测下一个即将出现的词语。它定义了哪些词能跟在上一个已经识别的词的后面（匹配是一个顺序的处理过程），这样就可以为匹配过程排除一些不可能的单词。

语言建模能够有效的结合汉语语法和语义的知识，描述词之间的内在关系，从而提高识别率，减少搜索范围。对训练文本数据库进行语法、语义分析，经过基于统计模型训练得到语言模型。

（5）语音解码和搜索算法

解码器是指语音技术中的识别过程。针对输入的语音信号，根据己经训练好的HMM声学模型、语言模型及字典建立一个识别网络，根据搜索算法在该网络中寻找最佳的一条路径，这个路径就是能够以最大概率输出该语音信号的词串，这样就确定这个语音样本所包含的文字了。所以，解码操作即指搜索算法，即在解码端通过搜索技术寻找最优词串的方法。

连续语音识别中的搜索，就是寻找一个词模型序列以描述输入语音信号，从而得到词解码序列。搜索所依据的是对公式中的声学模型打分和语言模型打分。在实际使用中，往往要依据经验给语言模型加上一个高权重，并设置一个长词惩罚分数。

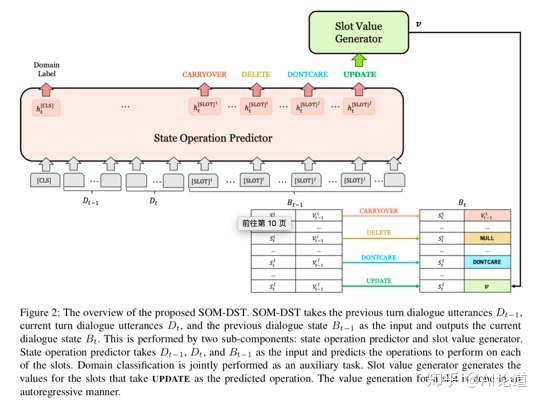
### 回答状态跟踪

下面我们介绍三篇对话状态追踪的相关论文，分别可以概括为检索式、生成式、生成检索结合，从而了解当前主流的DST模型。

01检索式

SUMBT: Slot-Utterance Matchingfor Universal and Scalable Belief Tracking

论文地址: [https://arxiv.org/abs/1907.07421](https://link.zhihu.com/?target=https%3A//arxiv.org/abs/1907.07421)



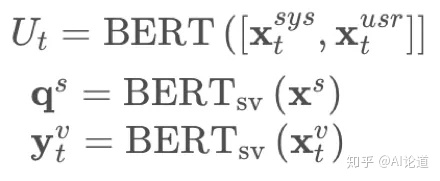
过去的方法，存在两个问题：

1）domain和slot分开建模，

2）实际场景中出现新的value值时，很难预测。本文将domain和slot一起编码，通过bert学习context和槽值对的内在联系，从而建立一个高效轻便的检索式模型。

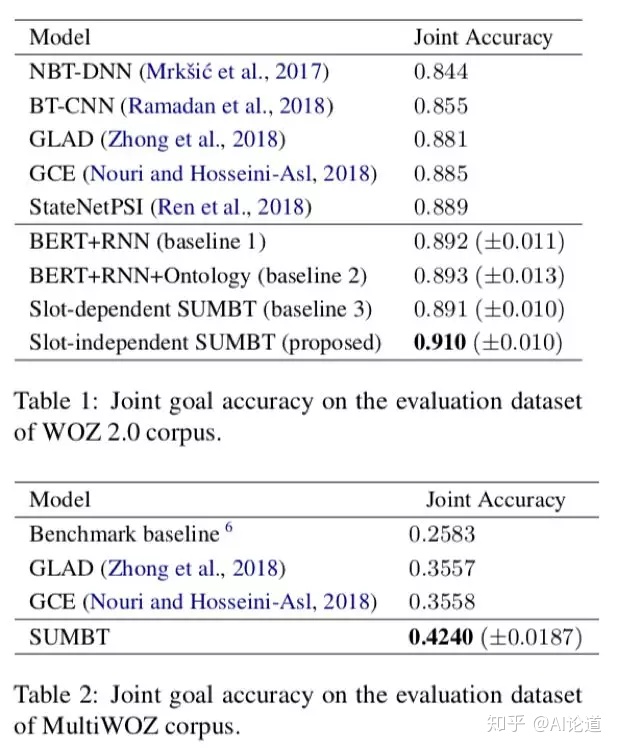
具体算法如下：

1）Bert对context，某个domain-slot以及其对应的所有value取值集合分别进行encode，均取encode输出的[CLS]对应位置为隐层表示。



2）d作为距离度量指标，用来判断domain-slot对应值是否为当前value，文中使用欧式距离和负向余弦距离(negative cosine distance)来计算。

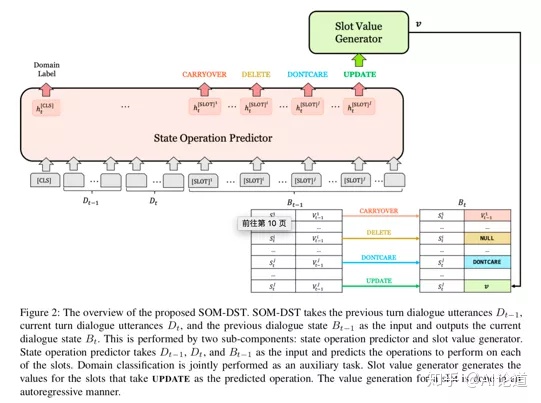
实验结果如下：



02生成式

Efficient Dialogue State Tracking by Selectively Overwriting Memory

论文地址: [https://arxiv.org/abs/1911.03906](https://link.zhihu.com/?target=https%3A//arxiv.org/abs/1911.03906)



过去的方法，每轮DST预测时，都需要重新输入已发生对话的所有语句，效率较低。本篇论文维护一个状态追踪表，并提出一个选择性重写机制。算法过程可以概括为：

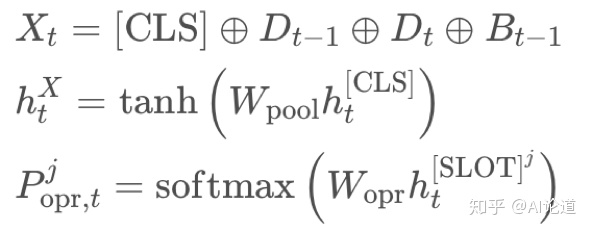
1）预测每个槽值对的状态操作，

2）对需要更新的槽值对进行重写。

状态操作（state operation）分为4类：CARRYOVER, DELETE, DONTCARE, UPDATE。CARRYOVER表示某个槽值对不变，剩余三个操作分别表示: 删除为NULL、槽值对未提及、槽值对更新。模型输入上轮回答、当前问句、上轮对话状态，只有当某个槽值对的状态操作为UPDATE时，模型生成对应的更新值。

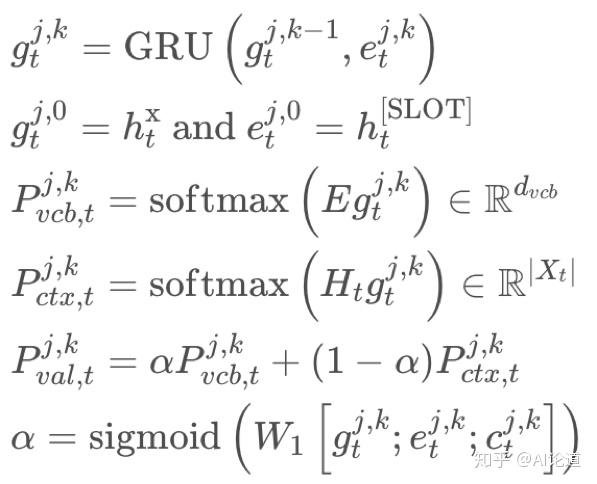
具体算法如下：

1)状态操作算法：

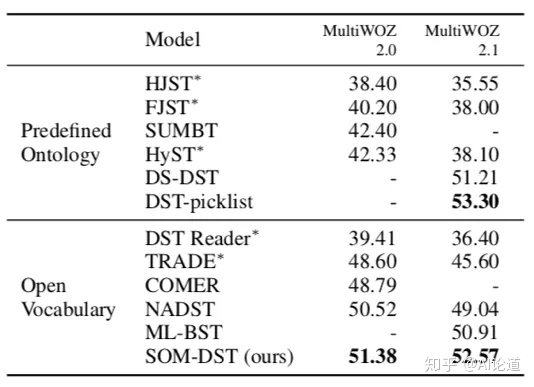


通过bert进行信息编码，输出的首位表示为domain信息，之后为各槽值对的状态操作预测值。

2)槽值对生成算法：



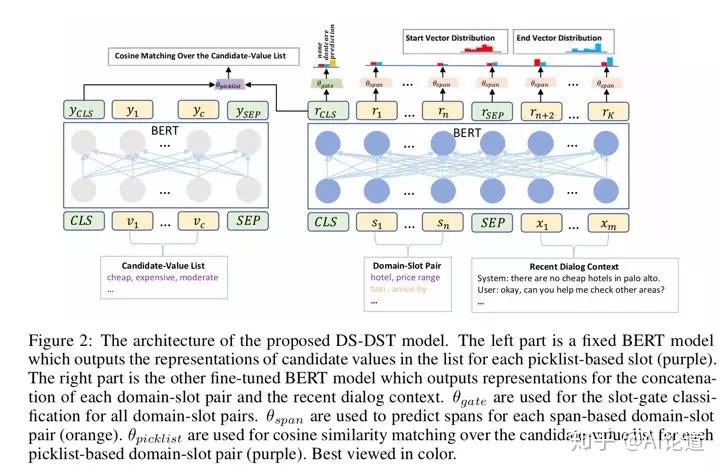
实验结果如下：



03检索生成结合

Find or Classify? Dual Strategy for Slot-Value Predictions on Multi-Domain Dialog State Tracking

论文地址: [https://arxiv.org/abs/1910.03544](https://link.zhihu.com/?target=https%3A//arxiv.org/abs/1910.03544)



目前DST大致分为两类，检索式(picklist-based)和生成式(span-based)。

检索式方法预定一个所有可能的槽值候选集，使用分类方法确定某个slot对应的value取值。生成式方法中，通过对context进行深度建模分析，找出语句中的value值。两种方法各有利弊，检索式方法高效不易错，但实际应用中构建完整的槽值候选集有较大的难度。生成式方法灵活且不受候选集依赖，但特定value(例如时间地点)较难抽取。

本文将两种方法结合，把所有slots分成两类，对于picklist-based slots使用predefined ontology-based approach，类似上文中的SUMBT ;而对于span-based slots, 使用span extraction-based method，找出某slot的对应value在context中的起止、终止位置。

具体算法如下：

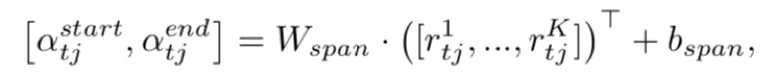
1）Bert编码输入context和domain-slot:



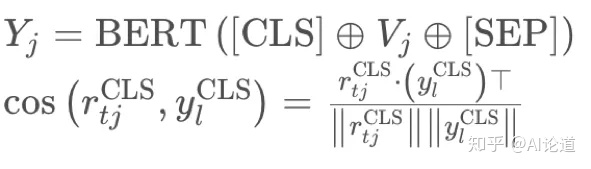
2）通过slot-gate分类算法判断一个domain-slot是否需要进行状态更新。(一个domain-lost可能对应的所有状态值，none, dontcare, prediction):



3）对于span-based slots:



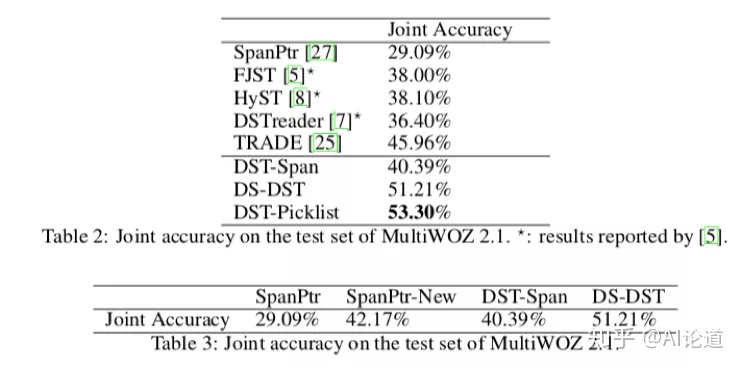
4）对于picklist-based slots:



训练时，将2)、3)、4)的loss函数相加作为模型loss。

论文在实验部分详细讨论两类slots的区分方式。比如Threshold-10表示candidate-value lists的长度少于10则划入picklist-based slots，否则划入span-based slots。然而实验表明，基于人工启发法进行区分时，模型性能更好。人工启发法是指把time和number这两种槽位划入span-based slots。

实验结果如下：



### 自然语言生成NLG

NLG 是为了跨越人类和机器之间的沟通鸿沟，将非语言格式的数据转换成人类可以理解的语言格式，如文章、报告等。

自然语言生成 – NLG 有2种方式：

1. text – to – text：文本到语言的生成
2. data – to – text ：数据到语言的生成

**NLG 的3个 Level**

简单的数据合并：自然语言处理的简化形式，这将允许将数据转换为文本（通过类似Excel的函数）。为了关联，以[邮件合并](https://baike.baidu.com/item/%E9%82%AE%E4%BB%B6%E5%90%88%E5%B9%B6/7804213?fr=aladdin)（MS Word mailmerge）为例，其中间隙填充了一些数据，这些数据是从另一个源（例如MS Excel中的表格）中检索的。

模板化的 NLG ：这种形式的NLG使用模板驱动模式来显示输出。以足球比赛得分板为例。数据动态地保持更改，并由预定义的业务规则集（如if / else循环语句）生成。

高级 NLG ：这种形式的自然语言生成就像人类一样。它理解意图，添加智能，考虑上下文，并将结果呈现在用户可以轻松阅读和理解的富有洞察力的叙述中。

**NLG 的6个步**

第一步：内容确定 – Content Determination

作为第一步，NLG 系统需要决定哪些信息应该包含在正在构建的文本中，哪些不应该包含。通常数据中包含的信息比最终传达的信息要多。

第二步：文本结构 – Text Structuring

确定需要传达哪些信息后，NLG 系统需要合理的组织文本的顺序。例如在报道一场篮球比赛时，会优先表达「什么时间」「什么地点」「哪2支球队」，然后再表达「比赛的概况」，最后表达「比赛的结局」。

第三步：句子聚合 – Sentence Aggregation

不是每一条信息都需要一个独立的句子来表达，将多个信息合并到一个句子里表达可能会更加流畅，也更易于阅读。

第四步：语法化 – Lexicalisation

当每一句的内容确定下来后，就可以将这些信息组织成自然语言了。这个步骤会在各种信息之间加一些连接词，看起来更像是一个完整的句子。

第五步：参考表达式生成 – Referring Expression Generation|REG

这个步骤跟语法化很相似，都是选择一些单词和短语来构成一个完整的句子。不过他跟语法化的本质区别在于“REG需要识别出内容的领域，然后使用该领域（而不是其他领域）的词汇”。

第六步：语言实现 – Linguistic Realisation

最后，当所有相关的单词和短语都已经确定时，需要将它们组合起来形成一个结构良好的完整句子。

**NLG 的3种典型应用**

NLG 的不管如何应用，大部分都是下面的3种目的：

1. 能够大规模的产生个性化内容
2. 帮助人类洞察数据，让数据更容易理解
3. 加速内容生产

下面给大家列一些比较典型的应用：

自动写新闻

某些领域的新是有比较明显的规则的，比如体育新闻。目前很多新闻已经借助 NLG 来完成了。

聊天机器人

大家了解聊天机器人都是从 Siri 开始的，最近几年又出现了智能音箱的热潮。

除了大家日常生活中很熟悉的领域，客服工作也正在被机器人替代，甚至一些电话客服也是机器人。

BI 的解读和报告生成

几乎各行各业都有自己的数据统计和分析工具。这些工具可以产生各式各样的图表，但是输出结论和观点还是需要依赖人。NLG 的一个很重要的应用就是解读这些数据，自动的输出结论和观点。（如下图所示）

**总结**

自然语言生成 – NLG 是 NLP 的重要组成部分，他的主要目的是降低人类和机器之间的沟通鸿沟，将非语言格式的数据转换成人类可以理解的语言格式。

NLG 的3个level:

1. 简单的数据合并
2. 模块化的 NLG
3. 高级 NLG

NLG 的6个步骤：

1. 内容确定 – Content Determination
2. 文本结构 – Text Structuring
3. 句子聚合 – Sentence Aggregation
4. 语法化 – Lexicalisation
5. 参考表达式生成 – Referring Expression Generation|REG
6. 语言实现 – Linguistic Realisation

NLG 应用的3个目的：

1. 能够大规模的产生个性化内容
2. 帮助人类洞察数据，让数据更容易理解
3. 加速内容生产

NLG 的3个典型应用：

1. 自动写新闻
2. 聊天机器人
3. BI 的解读和报告生成

### 语音合成TTS

一、核心概念  
1、TTS（Text-To-Speech，从文本到语音）

我们比较熟悉的ASR（Automatic Speech Recognition），是将声音转化为文字，可类比于人类的耳朵。

而TTS是将文字转化为声音（朗读出来），类比于人类的嘴巴。

大家在siri等各种语音助手中听到的声音，都是由TTS来生成的，并不是真人在说话。

TTS的实现方法，主要有2种：“拼接法”和“参数法”

2、拼接法

1）定义：从预先录制的大量语音中，选择所需的基本单位拼接而成。这样的单位可以是音节、音素等等；为了追求合成语音的连贯性，也常常使用双音子（从一个音素的中央到下一个音素的中央）作为单位。

2）优点：语音质量较高

3）缺点：数据库要求太大。一般需要几十个小时的成品预料。企业级商用的话，需要至少5万句，费用成本在几百万元。

3、参数法  
1）定义：根据统计模型来产生每时每刻的语音参数（包括基频、共振峰频率等），然后把这些参数转化为波形。

主要分为3个模块：前端处理、建模和声码器。

前端做的事情，是把文本进行解析，决定每个字的发音是什么，这句话用什么样的语气语调，用什么样的节奏来读，哪些地方是需要强调的重点，等等。常见的语气相关的数据描述包含但不限于下面这些：韵律边界，重音，边界调，甚至情感。 还有更多的信息甚至是难以客观描述的，目前的算法只能暂且忽略。

注：拼接法和参数法，都有前端模块，拼接和参数的区别主要是后端声学建模方法的区别。

2）优点：数据库要求相对较小一些。

如果只需要出声（做demo），大概500句就可以，但是效果肯定不行。

通用TTS，一般至少需要5000句，6个小时（一般录制800句话，需要1个小时）。——从前期的准备、找人、找录音场地、录制、数据筛选、标注，最终成为“可以用的数据”，可能至少需要3个月。（讯飞在各方面比较成熟，用时会短很多）。

个性化TTS，大多数是用“参数”方法的。（adobe、微软也有尝试过拼接法，不过相对参数方法来说不是太成熟，效果也并不是太通用）。

3）缺点：质量比拼接法差一些。因为受制于发声算法，有损失。

因为主要弱点和难点就是声码器。声码器的作用是复现声音信号，难在重现声音细节，并且让人听不出各种杂音、沉闷、机械感等等。目前常见的声码器都是对声音信号本身作各种理论模型以及简化假设，可以说对细节的描述近似于忽略。

而DeepMind的wavenet，基本解决了声码器的问题。因为他们直接对语音样本进行预测，不依赖任何发音理论模型。最后出来的音质细节十分丰富，基本达到了与原始语音类似的音质水准（所谓质量提高了50%，就是这里）。而且几乎可以对任意声音建模（这就太牛了）。目前wavenet的缺陷是计算量问题，速度太慢。大概100倍实时的量级，所以没有办法直接实用。

今年开始火起来的end-to-end的TTS建模方法，加上wavenent的声码器思想，才是未来TTS的发展方向。

4、TTS的评判标准

1）主观测试

A）MOS（Mean Opinion Scores），专家级评测（主观）；1-5分，5分最好。

微软小冰公开宣传是4.3分，但有业内朋友认为，也不能据此就说其“绝对”比科大讯飞好，因为每次评审的专家人选都不一样。说白了，目前整个AI行业内，还是各家说自己好的节奏。

B）ABX，普通用户评测（主观）。让用户来视听两个TTS系统，进行对比，看哪个好。

C）每次主观测评应该有区分。比如这次着重听多音字，下次主要听语气词等。

2）客观测试

A）对合成系统产生的声学参数进行评估，一般是计算欧式距离等(RMSE，LSD)

B）对合成系统工程上的测试：实时率(合成耗时/语音时长)，首包响应时间(用户发出请求到用户感知到的第一包到达时间)。

二、当前技术边界

1、通用TTS  
1）在用户预期不苛刻的场景（APP/硬件），能满足商业化需求，比如语音助手/滴滴/高德/智能音箱/机器人）；但如果用户预期非常高的话，是很难满足的，因为还是会有“机器感/机械感”，不能非常自然的模拟人声。

2）行业各家公司的产品效果差不多，都基本能商用；当然，讯飞会稍好一些。

2、个性化TTS

1）在用户预期不苛刻的场景，能“基本”满足商业化需求，但是效果没通用TTS那么好。但如果用户预期非常高的话，是满足不了的。

2）目前行业内能成熟商用的，主要还是科大讯飞。

3、情感TTS

1）目前业界的情感合成更多了，是因为数据本身变多了、更有节奏了，超过了传统的播音风格，但并不是真正的“喜怒哀乐”等情感合成（想高兴就高兴的这种智能）。

2）在情感TTS的理论方面，学术界是有储备的，但是，整个行业都没怎么做（或者没做好），是因为情感TTS很依赖“情感意图识别”，“情感特征挖掘”、“情感数据”以及“情感声学技术”等，是个系统工程。其中第1点，即是和自然语言处理相关，比如需要知道“什么时侯该高兴或悲伤”。

三、瓶颈和机会

1、基础技术

1）end-to-end的TTS建模方法，加上wavenent的声码器思想。（来源：<https://www.zhihu.com/people/meng-meng-55-13/answers> ）；

2）据报道，百度的Deep Voice团队（在美国硅谷的AI Lab），实时语音合成神经网络系统（Real-Time Neural Text-to-Speech for Production）比 WaveNet 要快 400 倍。

3）如何让离线版效果达到在线版水平。很多客户希望（奢望）有离线版本，并且效果和在线版本一样好。

2、数据缺乏  
一方面，特别是个性化TTS，需要数据量更大。比如默认男孩声音，要转成女孩，就比较难。另一方面，数据的获取（制作）成本和周期，也是各家在初期的竞争着力点。

3、人才匮乏：不仅没法跟NLP、CV等热门AI人才比，就算跟同样不算热门的ASR比，TTS的人才都还要少一些。

4、产品化难度：由于技术限制，现阶段不可能有非常完美的TTS效果，所以

1）尽量选择用户预期不苛刻的场景，或者在产品体验设计时，管理好用户预期。（比如打车软件，郭德纲/林志玲的声音，差不多就行）

2）选择“参数法”还是“拼接法”，和公司的技术储备、成本、以及产品目标相关。在垂直领域，现有的TTS技术（参数或者拼接）都可以针对产品做得很好。现在行业还没有太好的效果，很大原因是因为产品经理还没有深入介入，有很多细节的坑要踩~未来一定会有惊艳的产品出现。

3）体验细节设计，和一般互联网产品很不同，比如：

A）文案设计，非常重要；因为在语音交互场景，不能太长，用户没耐心和时间听完的。

B）可以加入背景音乐，掩盖杂音等细节瑕疵。

C）特殊场景，还有有特别的需求，比如远场TTS，和戴耳机场景，还会区别。

D）中英文混合TTS。比如用户想播首英语歌曲，困难在于：所有中文的发音当中，中文和英文合拍念出来是很难的，为什么呢？因为往往录音的人。录中文是一批人，录英文又是一批人。两种语言结合起来，再用机器学习学出来，声音就会变得非常怪。小雅音箱找到了一个能够和中文发音很像的女孩子，录了很多英语的音。

5、商业化压力

1）如果要有足够的市场竞争力，至少需要12个月的时间，2~6人团队，几百万资金投入（1个GPU一年十万，支持并发只有几十个）。并且，大公司的先发优势巨大，小公司必须切细分场景。

2）我个人认为，个性化TTS、情感TTS会在各细分场景得到更大的应用，比如知识付费、明星IP、智能硬件、机器人等。

### 问题解析：

处理问题：处理用户输入的自然语言问题，系统对于问题进行处理和分析，并对问题进行分类，确定问题类型；

生成搜索关键词：问题中的一些词不适合作为搜索关键词，另一些词的搜索权重则较高。系统需要对于用户的问题进行分析，来获得不同关键词的权重。

信息检索：系统使用从用户的问题中得到的关键词，对于数据库中的文档与关键词的计算匹配程度，从而获取若干个可能包含答案的候选文章，并且根据它们的相似度进行排序；

### 答案抽取：

段落提取：段落(paragraph)是包含答案的一个小节。问答系统与搜索引擎的区别在于用户期望其返回精确的答案，而不是一个文章或段落。为此首先要从文章中提取出可能包含答案的段落；

答案提取：在答案可能出现的段落被提取到以后，问答系统需要精确抽取段落中所包含的答案。这一步会用到问题分类。同时根据问题的关键词，对于段落中的词进行语义分析，最终找到最有可能是答案的字段。

## 2.4研究现状

根据上面的阐述可以看到，根据不同的技术路线，检索式问答、社区问答以及知识库问答所采用的评测数据集也不尽相同。

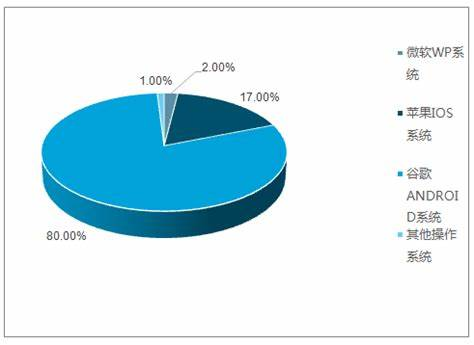
在检索式问答方面，最权威的评测是美国国家标准技术研究所（NIST）推动的TREC（TextRetrieval EvaluationConference）于1999年开始组织的问答评测任务（QA Track）8和NTCIR（NII Testbeds and Communityfor Information access Research）组织的跨语言问答评测任务（CLQA）9。TREC QA评测考察三类不同的问题：事实性（factoid）、列表类（list）和定义类（definition）。然后综合这三类问题的平均得分，对于参评系统进行评价。TREC QA 评测一直持续到 2007 年，该评测一直都是检索式问答领域最受关注、参加机构最多的 TREC 评测项目之一。根据 2007 年的评测结果来看，最好的评测系统 MRR（MeanReciprocal Rank）可以达到 0.48 以上（接近0.5意味着评测系统对于所有的问题将在前两位返回结果中获得正确答案）。基于 TREC 评测系统，IBM公司于2011年研发的 Watson 系统参加了美国 Jeopardy 知识比赛，并战胜了人类选手，可以看做是检索式问答系统的一个里程碑。但是不可忽略的是，Jeopardy比赛还是一个限定问题类型、限定答案类型的知识比赛，面对开放式的场景和环境，已有检索式问答系统还有很长的路要走。

在社区问答方面，目前并没有权威的评测数据集，公认的数据集通常是由 Yahoo!Answers 社区问答系统上利用为研究人员提供的 API10接口下载的。目前，最好的检索系统在 Top 10 的准确率可以达到 40%。尽管社区问答系统相对于检索式问答和知识库问答技术简单，但是目前已经商业化，例如 Yahoo Answer11和百度知道12。

在知识库问答方面，已有的评测主要针对于一些限定领域的知识库进行问答。已有方法也取得了不错的结果。例如：在Geoquery13（美国地理知识查询）数据集上（600个训练样本，280个测试样本）上，使用 CCG 和本体匹配的方法F值能达到89.0%，使用DCS的方法F 值能达到 91.1%；在求职（JOBS）数据集上（500 个训练样本，140 个测试样本），使用 CCG的方法F值能达到 79.3%,使用 DCS 的方法F值能达到 95%。在这一方面，QALD（QuestionAnswering overLinked Data）评测的举办更是推动了这方面的研究。QALD 每年举办一届，目前已经举办到了第六届。每一次评测，组织者都会给出一些问题，要求参加评测系统在给定知识库的基础上，将所给问题转化为结构化的SPARQL 查询语句，并在给定知识库上查询答案。但是，目前的研究趋势是从限定领域的知识库向大规模开放域甚至是多领域知识库进行扩展，例如 Freebase。与限定领域知识库相比，大规模开放知识库包含的资源和关系数量要大得多，比如Geoquery中只包含8个关系谓词，而 Freebase 包含上万个关系。因此开放知识库上的语义解析效果有明显下降。例如利用Freebase知识库，开放查询测试的最好效果只有39.9%；而在 QALD 评测中，在DBpedia上、开放查询中，表现最好的问答系统的正确率只有 40%。下图给出在面对开放域知识库 Freebase 时，在公开问题库 WebQuestion上，已有系统能够达到的精度。

## 2.5对比分析





### 苹果 Siri

### 微软小冰和小娜

微软针对娱乐化和工具化这两个截然不同的定位，分别推出了小冰和小娜（Cortana）。小娜，作为嵌入在 Windows 或 Windows Mobile 等微软操作系统内核的语音个人助理，承载着类似 Siri 或 Viv 的角色，它的目的是提升用户的工作效率，据说 Cortana 有 1.5 亿多用户，这也使得微软吸引到 Bengio 这样的大师作为顾问加入。另一方面，小冰是微软中国团队推出的娱乐聊天机器人。她的人设是一位 16 岁的少女。小冰是一个基于搜索的回复检索系统。通过各种基于深度学习的语义匹配算法，从海量的问答对语料中返回最佳的回复（Message response 而非 Answer）。小冰也会不定期推出新的技能供大家使用，这些技能往往包含了微软团队在图像理解、语音和自然语言理解方面的各种小应用尝试。更值得一提的是：微软针对日本、北美和欧洲等市场陆续推出了具有不同人设的少女如 Rinna、Tay和 Zo，她们往往可以方便的通过微信、微博或 Twitter 等平台进行交流。

### IBM Watson 系统

Watson 系统是典型的问答系统，其由 IBM 研究院在 2011 年推出，参加美国知识竞赛Jeopardy!（危险边缘）并挑落人类冠军而名声大躁。相比 AlphaGo 或早年 IBM 研制的战胜卡斯帕罗夫的国际象棋人工智能程序深蓝，Watson 具有更清晰的商业路径。IBM 斥巨资成立医疗事业部，并与 MD Anderson 等知名医疗机构合作推出面对特定病种（尤其是癌症）的辅助诊断 AI 医生。与此同时，Ross Intelligence 依托 Watson 认知计算平台推出了法律咨询系统。回到技术层面，Watson 所用到的知识库是一个广义的知识库，不仅包含各种结构化知识、也包含各种文本语料和语言学知识。整个流程称为 Deep QA，包含问题分解、假设生成、基于证据的融合排序等关键步骤。这里的 Deep QA 并非指通过深度学习（Deep Learning）技术来提供问答。事实上，Watson 诞生于深度学习大热之前，这里的Deep 是指通过深度解析（Deep Parsing）来实现对问句的真正理解。

### Facebook Messenger

### 亚马逊Alexa

### Google Assistant

从 Google Now 到 Google Assistant，谷歌一直没有停止过在语音个人助理方面的尝试。Allo是基于 Google Assistant 的新一代人工智能类微信 IM 应用。Allo 具有几个亮点：首先，其具备一定的自我学习能力。这里包括两方面的学习，一方面是学习用户的习惯，包括说话风格和交互模式。值得一提的是，Allo 的开发者也参与了 Gmail Smart Reply功能的开发，帮用户草拟回复的邮件。具体来说，根据邮件接收的对象、主题和关联的场景等，根据用户口吻来尽量完成要回复内容。另一方面也包括用户偏好的学习，这一点在推荐系统中是非常重要的，属于用户画像的学习。Allo 学习用户画像的低维稠密向量化表示（User Embedding）。将 User Embedding 加入 Chatbot 的回复生成解码模型中，将有助于回复的相对一致性和个性化。

### 百度度秘

### 搜狗汪仔

搜狗汪仔机器人是搜狗公司打造的问答机器人，搜狗汪仔能听、会说、会看、会思考，代表了人工智能领域的前沿技术。搜狗汪仔背后依靠的是基于人工智能技术的搜狗立知问答系统。立知问答系统是搜狗搜索研发的面向未来的搜索技术，在理解用户的问题或信息需求后，利用海量网络信息及大规模知识库，直接给出答案，方便快捷，适用于多种交互方式。其背后蕴含语义分析、问题理解、信息抽取、知识图谱、信息检索、深度学习等众多技术。

### 阿里小蜜

### 科大讯飞灵犀

灵犀是中国移动和科大讯飞联合推出的智能语音助手，更是国内首款支持粤语的语音助手。灵犀既能语音打电话、发短信、查天气、搜航班，还能查话费、查流量、买彩票、订彩铃，还可以陪你语音闲聊讲笑话。

### 其它

## 2.6未来展望

1)扩大适用人群：一方面采用简单、方便、快捷、精准的服务，使得一般用户都能够使用;另一方面提高客户服务系统的知识范围，能够满足更大范围的人群

2)统一服务接口：当前的大多数客户服务系统提供多个服务接口，如售前接口、售后接口、咨询接口等，若能够提供统一的接口，这样用户不需要进行选择或切换，就可以使用统一的接口享受服务。

3)更好的语义性：在线问答客服系统通过自然语言进行问答，系统能够很好的理解用户的上下文语境和语义，做出相关的应答，当前优秀的问答系统虽然能够对单个语句的语义信息进行理解，但是上下文信息还是没有得到很好的处理;

4)语音识别：传统的问答客服系统的问答方式一般为文本输入以及文本输出，是将网络上大量的相关信息而不是准确信息提供给用户，供用户自己进一步去筛选。基于语音识别、智能搜索的语音问答客服系统，能够根据用户的语音自动搜索出结果，并且将结果以语音返回给用户。

5)更强大的知识库：无论是线上问答客服系统还是线下的电话客服中心，普遍存在知识匮乏的问题，构建一个强大的知识库满足客户的需求尤为重要。

6)智能化：随着物联网和移动互联网的发展，智能终端越来越为人们更多的接触，通过手机终端、电视终端等，以友好的接口以文字或语音进行提问，系统经过处理返回相关答案。

7)社区化：以灵活交互的特点，满足客户获取和分享知识的需求。客户将其信息需求以问题的方式提交给系统，并等待其他用户给出答案。其他客户或客服人员根据个人兴趣、知识水平，选取适当的未解决问题来回答，实现知识的共享;然后客户对得到的答案给出评价并选择最满意的一个作为最佳答案。

8)个性化：结合用户信息和用户在客服平台的问答信息、使用行为、习惯、偏好和特点，挖掘用户感兴趣的问题，以个性化方式向用户提供服务。