对话智能与认知型口语交互界面

俞 凯
上海交通大学

关键词：对话智能 任务型对话系统 认知计算

随着互联网的形态从PC互联网向手机无线网和硬件物联网发展，人与智能设备的信息交互模式也发生了颠覆性改变。首先是输入模态的变化。传统的键盘鼠标正在被触感、语音、图像等模态的自然输入方式逐渐替代。从复杂语义信息传输的角度看，语音具有得天独厚的优势，是未来人机信息交互的核心入口。其次是用户获取信息习惯的变化。移动环境下的信息搜索行为呈现碎片化、本地化、个性化、情境化等特点，信息获取的目的性更强，使得基于理解的“对话交互式”信息获取成为新兴的语言处理关键技术。这两个趋势也使得以语音为主要通道的“对话智能”成为人工智能研究的新兴问题。

对话智能是人工智能技术的集中体现，目前尚无公认可行的通用处理框架理论。从语义结构化和交互方式角度来说，人机对话大体有三种主要类型。

1. 问答。这类对话一般有两大特点，一是单轮为主，对话模式是一问一答，不涉及复杂的对话上下文；二是语义非结构化，很难用预定义的本体和语义槽来表达整个对话任务。问答往往涉及到非结构化的知识表达、搜索以及在回答中的匹配。问答的目的在于完成任务和提取知识点，有非常明确的信息需求。

2. 聊天。聊天一般是娱乐或消磨时间，没有实用性的明确目的。这类对话通常是多轮交互，上下文相关，语义也是非结构化的。

3. 任务型对话。任务型对话系统针对具体的应用领域，具有比较清晰的业务语义单元定义、本体结构以及用户目标范畴，例如航班信息查询、电影搜索、设备控制等等。这类交互是以完成特定的操作任务为交互目标。此外，任务型对话绝大部分都是多轮的，需要结合对话上下文进行用户意图理解。

三种对话系统的应用场景、工程架构及核心技术不尽相同。“对话智能”本质上应该是上述三种类型对话的有机组合。问答和聊天式的对话在传统自然语言处理领域已有较长时间的研究，而任务型对话系统，尤其是口语对话系统的理论与技术，随着移动互联网和物联网的发展，开始得到学术界和产业界的重视，是本文讨论的重点。

典型的任务型口语对话系统的架构如图1所示。

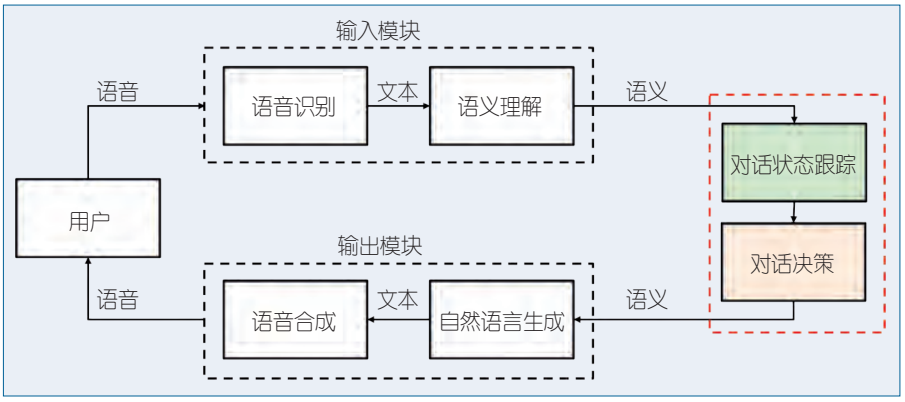

图1 任务型口语对话系统架构图

对话智能中的认知计算

任务型口语对话技术有时会被误解为传统语音识别与自然语言处理技术的简单组合，而实际情况并非如此。传统语音识别与自然语言处理技术的研究都没有覆盖两个新的、独特的认知问题，一是“对话管理”，即图 1 所示的“对话状态跟踪”和“对话决策”，这也是多轮交互特性产生的序列决策问题；二是由于语音识别无法避免错误以及用户自然交互的语义随意性大，口语交互中的语义理解结果往往具有非精确性和非连贯性，这使得“基于非确定信息的推理”成为必须要解决的重要鲁棒性问题。它要求机器的角色从简单的感知命令的“执行主体”变成可以和用户深度沟通的“认知主体”，实现认知型口语交互界面。

任务型对话技术是基于理解的对话技术，一般需要有语义信息的结构化表达。与传统自然语言处理中的语义框架不同，基于预定义好的语义本体的“对话动作”更为常用，它侧重于从行为的角度表示对话语义。1999 年，特劳姆 (Traum) 发展了对话系统中“行为”的概念，考虑了对话的轮次信息以及用行为来表达对话的意义，其中包括请求确认 (confirm)、询问 (request)、肯定 (affirm)、否定 (deny) 等。例如，“你是明天上午 10 点出发吗？”这是一种请求确认行为；而“你从哪里出发？”则是一种询问行为。为了表达更为具体的意思，对话行为通常会与语义槽值对 (slot-value pair) 组合在一起，形成现在广为使用的简单的“对话动作”形式 (act_type(a=x, b=y))。其中，act_type 是对话行为，a=x 和 b=y 则表示语义动作涉及的语义槽值对（如“出发时间 = 上午 10 点”）。更简单的语义槽值对有两种，语义槽为空和值为空，如 request(电话) 可以表示“电话是多少”，inform(=don’t care) 可以表示“我无所谓”。

如果已经有语义理解的算法可以从文字输入中解析出对话动作，则对话管理的本质就是机器根据用户输入的对话动作，以完成任务为目标，生成机器对话动作的过程。这个过程就自然形成了一个对话动作的决策序列。假设我们要设计一个购票咨询的对话系统，典型的例子如下（见表 1）。

表1　任务型对话与对话动作的例子

角色	文字	对话动作
机器	请问您想购买哪里出发的机票?	request(出发城市)
用户	上海。	inform(=上海)
机器	是虹桥机场还是浦东机场?	select(机场=虹桥, 机场=浦东)
用户	浦东机场，10点的航班。	inform(机场=浦东，时间=10点)
机器	您是说明天上午10点吗?	confirm(日期=明天，时间=10点)
用户	是的。	affirm()

上述例子中，如果忽略语音识别和语义理解的误差，整个对话过程就被抽象为机器和用户的一系列交替出现的对话动作序列，而对话管理就是要根据序列的历史以及最近一次用户的输入去决策应该采取何种具体的对话行为。很显然，这样的决策过程无法用传统的有监督学习模型来描述，因为每一轮的对话动作选择是没有“标准答案”的。决策过程应当作为一个整体来衡量其性能，而决策过程的性能衡量，也就是整体对话系统的性能评估，需要有客观的指标才能使得机器的自动优化成为可能。最直观也最简单的评估方式是以“任务是否达成”作为指标，这样的二元判断适用性广，标准清晰易操作，在很多场景中可以通过制定规则实现自动对话评估。但由于只关注对话的结果，忽略了对话的过程和细节，该方法无法较好地顾及用户体验。一个简单的折衷是将指标由“是 / 否”变为分数，对任务是否达成给以不同的奖励值，同时，对每一轮对话给一个惩罚值（例如 -1），使得轮数越多惩罚越大。对整个多轮对话，采用奖励和惩罚值之和作为指标就能避免用户对冗长对话的体验不佳。

显然，人机对话的过程与人机下棋博弈的过程有类似之处，即用户和机器（类似博弈双方）轮流给出对话行为，最终由“是否达成任务”来判断机器决策过程的好坏（类似赢棋）。当我们试图用机器学习来优化决策过程时，很自然地也会采用与优化

人机博弈类似的方法，如强化学习。强化学习的一类经典的数学框架是马尔可夫决策过程 (Markov Decision Process, MDP)。它是一个五元组 $(S, A, P(\cdot,\cdot), R(\cdot,\cdot),\gamma)$。其基本原理是为对话决策过程定义一个有限的状态空间 S，用“状态”来描述各类可能的用户意图和对话历史组合，在状态空间上建立一个状态转移模型 $P_a(s_{t+1},s_t)=P_r(s_{t+1}|s_t,a)$ 来描述状态之间转换的概率关系，对话状态的估计被称为“对话状态跟踪”；同时定义一个有限的对话行为空间 A，描述各类可能的机器反馈；定义一个收益函数 $R_a(s_{t+1},s_t)$，在每轮机器给出对话行为并得到用户再次输入，促使系统转入下一个状态后，给出一个收益值（如前述的惩罚值），在整个对话结束后再给出一个总体性能收益（如任务是否达成），所有收益值以折扣系数 γ 作为权重进行加权累积得到最终收益。状态空间到对话行为空间的映射函数反映了具体每轮的决策规律，称之为“对话策略”。对话状态跟踪模型和对话策略模型可以是基于规则的，也可以是数据驱动的。机器学习研究的热点是数据驱动的模型，其优化过程就是调整“对话状态跟踪”和“对话策略”的参数，使得最终收益最大。强化学习的相关算法在萨顿 (Sutton) 的经典教材 *Reinforcement learning: an introduction* 中有详细介绍。

对话管理中的序列决策研究将强化学习引入自然语言处理，形成一类新兴的研究方向。而语音输入对于对话交互的影响则引发了另一类新的研究问题——基于非精确信息的理解和决策。例如，机器需要根据用户的语音指令去操作股票的买卖，语音输入仅有两个字：“买”或者“卖”。由于语音识别可能产生错误，我们需要某种形式的容错决策机制（如不进行任何操作）来防止意外买卖。首先考虑用经典的方法来实现这样的机制。假设“买”和“卖”的语音识别只通过声调特征来实现，经典方法需要两个阶段：先根据用户的声调将输入识别为“买”（三声）或者“卖”（四声），再将识别结果输入对话反馈决策模块产生操作行为。其中首要问题是分类，我们的目标是根据声调特征来对孤立字 char={买，卖}分类。由于“买卖”两个字发音相同，音调相近，分类错误不可避免，通常我们会对每个字（买/卖）估计一个声调特征的条件概率分布 P(tone|char)，如图 2 所示。

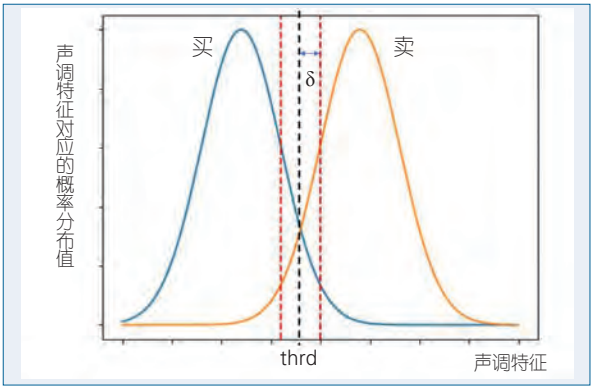

图2 买卖命令的声调分布

最优的决策边界可以根据后验概率 P(char|tone) 来确定。检测出声调特征后，可以通过与决策逻辑中的阈值进行比较来选择适当的实际购买操作行为。更进一步，可以根据后验概率计算决策错误率的概率分布，同时估计一个置信边界 δ。这样，经典方法的第二阶段就可以用一个简单规则或流程图来表示（如图 3）。

```
1：令 tone 为声调特征，thrd 为阈值门限，δ 为置信边界
2：if  tone < thrd - δ  then
3：    买股票
4：else if tone > thrd +  δ  then
5：    卖股票
6：else
7：    什么都不干
8：endif
```

图3 基于声调的股票买卖决策系统

那么，当采用这种经典决策机制时，我们丢失了什么？首先，没有明确的描述不确定性的模型。在上述例子中，虽然可以用置信边界来辅助决策，但是声调识别过程本身仍然是输出一个确定性的结果，而且一旦这个结果被后面的决策过程采用，就无法轻易去除影响。其次，没有尝试跟踪用户的意

图，系统无法确定它对声调的识别与用户的真实意图是否一致。系统可能会观察到用户很少在买入股票后马上就卖出同一支股票，但却很可能连续买入同一支股票，而这种行为特征（或针对意图的先验知识）完全可以用来消除声调识别不准确而产生的歧义。再次，由于没有量化指标，很难对流程图中的决策规则进行优化改进。这一切的后果是系统无法解释不确定性输入的含义，无法在不确定的条件下有效地进行规划，决策不具有在线适应变化的能力，也无法从经验中学习策略的改变。这些使得对话交互的界面不具有“认知能力”。

建立“认知型口语交互界面”的关键，是认识到“不确定性”（或非精确性、不准确性）是自然人机对话的本质属性之一。语音识别本身由于发音相近、噪声干扰、说话人语速、口音等问题具有不可避免的错误。多通道输入情况下，各个通道都有干扰产生的不确定性。另一方面，对于一般的对话交互，从认知角度，人类也自然地倾向于用非精确的信息进行交流，因为信息传输的速度能大大增加。因此，具有认知能力的口语对话系统，不将语音识别结果（或基于单一识别结果的语义理解结果）视为确定性的证据，而把它们视为一种特殊的观察特征，系统利用这些观察特征推断用户的所有可能意图。可以通过一组收益值来量化系统响应用户意图的有效性，必要的决策逻辑也可以通过最大化这些收益值而达到最优。

最初，上述方法的工程实现依赖于贝叶斯推理和贝尔曼最优化原则，而这个标准框架就是“部分可观测马尔可夫决策过程”(Partially Observable Markov Decision Process, POMDP)。它的数学描述是一个八元组 $(S, A, P(\cdot,\cdot), R(\cdot,\cdot), \gamma, O, Z, b_0)$，与五元组的马尔可夫决策过程不同的是，用户的实际意图，即状态 S 是不可直接观测的，可以直接观测到的是声调特征，用 O 表示。由于 S 是隐变量，我们只能通过状态转移函数和观测特征进行概率推断，其中 Z 定义了基于状态和机器行为的特征转移概率 $Z=P(o_t|s_t,a_{t-1})$。基于此，用来确定机器对话行为的“策略”函数，也就不再是“状态→机器行为”的映射，而是“状态分布→机器行为”的映射，b_0 表示初始的状态分布。杨 (Young) 在 *IEEE Signal Processing Magazine* 上的“Cognitive User Interfaces”中用类似例子给出了 POMDP 计算过程的详细说明。

近年来，深度强化学习因为 AlphaGo 而受到极大关注，它也正在被越来越广泛地用于统计对话管理的优化。虽然 POMDP 经典数学框架正被诸如深度 Q 网络、循环 Q 网络等深度学习方法代替，但其处理不确定性和进行策略估计的基本原理都没有变化。

现实中对话智能实现的挑战

上述两个简单例子说明了口语对话技术是以整体决策过程性能优化的思路，来处理非精确输入信息、语音和语言处理领域的一个新兴研究方向。虽然强化学习的基本理论用于任务型对话系统已经被学术界广泛接受，但对话智能并未达到实用水平，除了自然口语语义理解水平这个典型问题之外，在真实世界中，实现对话智能还有许多亟待解决的难题。

首先是对话状态和机器对话行为空间的规模问题。真实任务型对话系统的“状态”一般包括三个部分：用户意图、当前用户语义以及对话历史。对话历史可以简单地用任务语义槽的填充情况来表示（例如，每个语义槽定义“已确认”“未谈及”“正在进行”三个简单的可能取值），而前两部分通常以对话动作的方式表达。假设对话任务有 n 种对话动作，m 个语义槽，平均每个语义槽有 p 种取值，则一次用户输入的对话动作可能会有 np^m 种可能。以表 1 中的简单机票查询为例，如果只考虑 inform 一种对话行为，以及出发城市、到达城市、时间和日期 4 个语义槽，假定平均取值有 50 种，则每轮用户输入语义的种类可能有 600 多万个，再考虑用户意图及对话历史，这个状态空间的尺度将达到千万以上。与对话状态类似，机器可能产生的机器行为的空间也是巨大的。这就需要用一些结构化的方法来压缩对话状态空间和机器行为空间。一些语义本体结构的空间聚类压缩方法，如摘要空间算法等已

被广泛采用，其有效性已在若干小规模实验室级的真实系统上得到验证，但对于工业级的真实对话系统，如何有效地处理大尺度的对话状态和机器行为空间的描述仍然任重道远。

其次是对话管理模块的测试评估与用户仿真问题。客观量化的测试评估指标是进行数据驱动的对话管理优化的前提。但前文提到的对话任务完成度指标还是一个实验室条件下的度量，真实使用的对话系统往往得不到用户的明确反馈，而且由于口语对话系统是集成了识别、理解、合成等在内的综合体，即使得到用户反馈，也很难确定这些反馈是否都是针对对话管理的评估。研究者试图采用设计用户模拟器的方式来解决这个问题。用户模拟器精确了解用户意图等信息，通过它与对话管理器的直接交互（以对话动作为接口标准），可以有效地优化对话管理器。这一思路类似 AlphaGo 中的自我对弈。但是，它需要对用户交互习惯进行建模，虽然也产生了如议程模拟等一系列算法，但由于语言的复杂性，用户意图模拟的难度依然很大，目前仍是对话技术研究中的难点。一般来说，用户模拟器也仅仅用于初始化对话管理的参数，使其达到基本的性能水平，之后仍然期望在与真实用户的交互中持续优化。

上述问题同时引发了统计对话管理系统的冷启动困境。一方面，统计对话管理的优化需要通过与真实用户的交互才能实现；另一方面，系统要上线进行真实交互的前提必须是其性能达到较高水平，否则会因为影响用户体验而无法上线。这一困境也是目前数据驱动的对话管理系统尚未得到广泛使用的根本原因之一。最近，产生了一种新型的“混合智能”的解决思路，即采用人机混合提供对话服务的方式，在冷启动阶段借助人的交互来保持用户体验，同时“手把手”地优化机器策略，直至机器逐渐替代人来提供完整的对话服务。这也引发了一系列新的科学问题。

目前对话管理的理论研究中，主要关注的问题是如何对用户的输入进行响应。而在真正智能的对话系统中，系统应当有能力主动发起诱导性对话，预测用户可能的需求，给予提前响应；或者根据用户画像及对话历史，给出超出响应式回复的主动推荐。这类机器的主动对话能力是对话智能的高级体现，在真实的人工客服系统中屡见不鲜，但尚无基于数据驱动的理论框架来解决真实场景下的主动诱导对话问题。

上述任务型对话系统都是假定对话任务的语义本体是可以用语义槽的方式描述并预先定义的，而现实世界中，特定对话任务的语义本体可能很难用预定义的语义槽描述。这就涉及任务型对话中的语义表达与知识图谱等其他形态的知识（语义）表达之间的关系。面向交互的语义或知识表达，不仅需要一个有效的理论框架，还需要有动态扩展能力，因为“语义概念学习”本身也是对话系统认知能力的重要体现。同时，任务型对话与知识问答、聊天等对话形态的自然融合也需要语义及知识表达方面的理论进展。

在真实场景中，本文讨论的以文本语义为基础的轮回交替的对话方式也并不智能。从对话架构角度，用户与机器并不一定都是交替发言，而很可能出现连续发言；若干小段对话主题的发起、切换和结束，是真实对话中可能出现的情况；口语对话中的打断、容错、纠错等是自然口语交互过程的普遍现象。从模态角度，口语对话中传递信息的不只是语言，还可能包括各种副语言信息，如情绪、说话方式、交流连贯性，以及图像、触感等其他多模态输入。如何将多种信息融入基于语义理解的状态空间并进行有效推理，是对话智能的另一类挑战。

综上，对话智能是以“交互控制”和“非精确信息处理”为核心特点的新兴研究问题，又与知识及语义表达有直接的联系，在语音和自然语言处理领域中具有特殊的地位。目前，面向真实任务的口语对话系统大多以规则为主要方式在运行，对话智能的研究还有大量的处女地带，既有重大的理论挑战又有急迫的现实需求。 ■

俞　凯

CCF专业会员。上海交通大学研究员。主要研究方向为智能语音及语言处理和人机交互。

Kai.yu@sjtu.edu.cn