# 任务型人机对话系统中的认知技术
# —概念、进展及其未来

俞凯 [1,2], 陈露 [1,2], 陈博 [1,2], 孙锴 [1,2], 朱苏 [1,2]

[1](上海交通大学，计算机科学与工程系智能语音实验室，上海市，200240)

[2](上海市教委智能交互与认知工程重点实验室，上海市，200240)

**摘 要** 人机对话系统是将机器视为一个认知主体的人机交互系统。随着计算机软硬件技术和移动互联网的迅猛发展，能够有效处理非精确信息交互的，符合人类自然交互习惯的认知型人机对话系统受到了越来越多的关注。本文提出，面向任务的认知型人机对话系统的架构应分为三个层次：物理层、控制层和应用层，与之对应的技术是通道技术、认知技术和知识管理技术。其中，认知技术是随着移动实时交互的新的需求而产生出来的新的交互中间件技术。它的目标是使得机器具有人类的认知交互特点，可以在与对方的交互中进行深度理解、学习、诱导和适应，主要包括非精确信息理解技术、基于不确定性的推理和决策技术、交互自适应和进化技术，诱导式信息生成技术等。本文对认知技术在人机对话系统中的定位和具体概念进行了详细介绍，综述了相关技术领域的进展，并展望了未来重点的研究方向。

关键词 人机交互；认知技术；对话系统；人机界面；认知控制

**中图法分类号** TB391

# Cognitive Technology in Task-Oriented Dialogue Systems – Concepts, Advances and Future

Yu Kai [1,2], Chen Lu [1,2], Chen Bo [1,2], Sun Kai [1,2], Zhu Su [1,2]

[1]( SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

[2]( Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai, China)

**Abstract** Human-machine dialogue system is a human-machine interaction system which treats the machine as a cognitive agent. With the advances of computing hardware and software as well as the booming of mobile internet, cognitive dialogue system, which can deal with uncertain interactive information, attracts great interest. The paper argues that task-oriented dialogue system consists of three layers: physical layer, control layer and application layer. IO technology, cognitive technology and knowledge management are corresponding techniques. Cognitive technology is a new middle ware technology recently emerging with the need of instant natural human-machine conversation. Its goal is to make machine a cognitive agent capable of understanding, learning, guiding and adapting. For this purpose, deep and robust understanding, inference based on uncertain information, policy optimization, adaptation and influential information generation are required. This paper is a position paper of cognitive technology. The scope and content of cognitive technology in dialogue systems is introduced. Relevant techniques are reviewed and future research direction are also discussed.

**Key words** human-machine interaction; HCI; dialogue systems; cognitive technology; human-machine

---

本课题得到国家自然科学基金委优秀青年科学基金项目（No.61222208）；上海高校特聘教授（东方学者）岗位计划资助. 俞凯，男，1976年生，博士，研究员，中国计算机学会会员(E200030540M)，主要研究领域为认知型对话系统、语音合成、识别、理解、机器学习等，E-mail:kai.yu@sjtu.edu.cn. 陈露，男，1990年生，博士研究生，主要研究领域为统计对话系统、机器学习，E-mail:chenlusz@sjtu.edu.cn.陈博，男，1990年生，硕士研究生，主要研究领域为情感交互、机器学习，E-mail:bobmilk@sjtu.edu.cn.孙锴，男，1992年生，本科生，主要研究领域为人机交互、机器学习、人机对弈，E-mail: accreator@sjtu.edu.cn.朱苏，男，1990年生，硕士研究生，主要研究领域为语义理解、自然语言处理、机器学习，E-mail: paul2204@sjtu.edu.cn。

interface; cognitive control

## 1 引言

人机交互（Human Computer Interaction，HCI）是计算机诞生以来产生的研究人和计算设备之间相互影响的技术。其目标是机器帮助人高效、舒适、安全的完成任务需求。人机交互作为信息时代对人类生产生活具有重大影响的基础技术，受到广泛重视。美国在 2000 年制定的信息技术研究预算中，把“人机交互”与“软件”、“网络”和“高性能计算” 并列为四项基础研究[1]。对话系统（Dialogue System）是人机交互技术最核心的领域之一，它是人与机器之间进行双向信息交换以满足人的特定任务需求的计算机软硬件系统。广义上，对话系统包括所有人机交互系统，例如图形界面（Graphic User Interface，GUI）、虚拟现实交互等等。狭义上，对话系统特指完成类似人与人的沟通任务的计算系统，它的目标是使人机对话像人人对话一样有效、快捷和自然。这类交互系统强调机器在任务完成、信息交换和环境感知方面的拟人特性，把机器作为双向信息交互中的一个“认知主体”，因而“认知能力和相关技术”就成为此类系统的关键能力和特性。本文所讨论的就是狭义人机对话系统中的认知技术。

自上个世纪 20 年代，能说话的机器狗“Radio Rex”出现以来，智能的人机口语对话系统就成为人类梦寐以求的科技梦想。美国国防部高级研究计划署 (DARPA)非常重视推进大词汇连续语音识别技术，陆续设立了 EARS（2002-2005）、GALE（2006 - 2010）等数千万美金的大型研究计划；欧洲也投入巨资进行语音识别和合成的研究，例如英国的 NST（2011-2015）和欧盟的 AMI（2004-2008）、EMIME（2008 - 2012）计划。除识别和合成的研究之外，各国开始越来越重视理解和整体口语对话系统的研究，美国设立 CALO 计划（2003-2008），欧洲设立 TALK（2004-2006）、CLASSiC（2008-2011）和 PARLANCE（2011-2014）等项目用于资助对话系统相关技术的研究。产业界也不甘落后，Apple 公司购买了 CALO 计划的衍生公司，在 iPhone4S 上推出了著名的 Siri，谷歌公司在 2012 年提出，在移动互联时代，传统搜索要向语音对话式搜索转变。

这些趋势都显示了人机对话系统的重要意义。本文将对人机对话系统的框架及相关技术研究进行回顾和分析，提出自然人机对话系统的技术架构层次分类，并着重对其中的认知技术进行综合论述。

## 2 对话系统及其技术分层

人机对话系统是将机器视为一个认知主体的人机双向信息交换系统。最初的人机交互系统都是将机器看做是执行精确命令，产生预定的输入输出的工具。如命令行交互终端、图形用户界面和键盘鼠标交互等等。这类人机交互系统大都是以设计者为中心，要求用户按照设计者预定的方式进行交互并获取结果。而随着技术和应用的发展，以用户为中心的人机交互系统从上个世纪末开始受到越来越大的重视。这类交互系统不是要求用户去适应机器（交互系统设计者），而是要求机器去适应人，也即允许用户采用与人交流的自然方式去与机器交流。这就产生了一个观念上的变革，机器的角色从“执行主体”变成了“认知主体”，可以和人进行“对话”沟通。这类人机交互系统就是本文所关心的“人机对话系统”。

对话系统从本体构成和业务逻辑角度，可分为领域任务型和开放型的信息交互。领域任务型系统针对具体应用领域，具有比较清晰的业务语义单元的定义、本体结构以及用户目标范畴，例如航班查询、视频搜索、设备控制等等，这类交互往往是以完成特定的操作任务作为交互目标；而开放型信息交互则不针对特定领域，或说面向非常广泛的领域，交互目标并非业务任务，而是满足用户其它方面的需求，例如开放的百科问答、聊天等。它虽然能一定程度上显示人工智能的力量，但因其并不专注于帮助人解决现实任务问题，其实际使用范围较为狭窄。近年来，随着移动终端的高速发展，面向任务的自然人机对话系统和相关的认知控制理论得到了越来越多的学术和产业界重视，也是本文讨论的重点。

### 2.1 任务型对话系统的三个层次

领域任务型人机对话系统是一个闭环的双向连续信息交换系统，传统观点往往把它粗略的分为输入、理解和输出三个模块，其中“输入输出”和“理解”之间进行的是确定信息的交换，“理解”被笼统的认为是采用自然语言处理技术对输入文本进

行分析。这种分割方式忽视了信息的不确定性和人脑对各个模块进行整体的、系统的调度协调的认知能力，也没有把信息内容本身的管理与信息的调度和控制区分开来，而这些恰恰是现实的自然人机对话中不可回避的重要问题。因而，本文将任务型对话系统明确分为三个层次，以便讨论认知技术的范畴及接口。如图 1 所示：

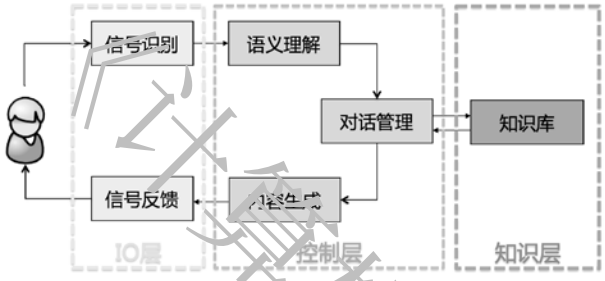


图 1 领域任务型的人机对话系统框图

- IO 层

最外部的输入输出层（IO 层）是对物理层面信号的处理，也即传统人机交互理论中的“通道”，其目标是对用户和环境产生的信号进行感知和通道层的分析，转换为一定的编码，它对应的技术范畴是通道感知技术。

- 知识层

最内部的知识层是对领域任务相关知识的管理，目标是对特定的知识进行对话前的离线预处理，如获取、搜索、索引等，这一层对应的技术范畴是知识管理技术。

- 控制层

控制层对 IO 层得到的编码进行语义解释，维护对话系统的认知状态空间，管理知识的交互式提取和交换，并进行对话推理和决策，它是 IO 层与知识层的中间件，这一层对应的技术范畴是认知控制技术。

知识层在领域任务型的对话系统中，通常以知识数据库的方式出现，可以进行精确或模糊的查询。这种知识管理的方式相对成熟和简单，知识数据库查询本身不是本文讨论的重点 [1]。控制层在确定语义本体和业务逻辑的前提下，与知识层是相对独立的，这样就使得控制层的研究可以独立于知识管理，形成完整的研究体系。控制层与IO层都与交互能力直接相关，它们紧密耦合在一起，就是“认知型人机界面”，是传统人机界面的扩展。值得指出的是，面向业务的口语问答系统也是一种特殊的任务型口语对话系统，它的知识层可能采用与传统的数据库查询不同的表达和管理方式（例如采用关键字匹配）。这类口语对话系统的认知技术与数据库查询型的对话系统的认知技术原理大部分是相似的，但在语义接口定义和对话状态表达上差别较大，更偏向于开放型信息交互（除非能转化为数据库查询型的知识管理方式），因此相关的细节不在本文中讨论。

## 2.2 通道技术

通道（modality）的概念往往指用户的行为或通信方式，通道技术是将信号转换为编码的技术。作为“对话”的天然载体，语音信号转换为文本编码一直是人机对话系统的主要通道形式。语音通道的核心是语音识别（输入）和合成（输出）技术。

对话系统中的语音识别通常是非特定人、非特定环境的大词汇连续语音识别，这类语音识别系统涉及特征提取、声学模型、语言模型和解码算法等关键技术，其目标是找到一个文本序列，使得它在给定音频上的后验概率最大。这个过程中，声学系统负责将声音转成基础发音单元（音素），而语言系统则运用语言知识纠正声学系统的识别错误。目前的商业主流技术是采用隐马尔可夫模型（Hidden Markov Model，HMM）结合深度神经网络（Deep Neural Network，DNN）[2]作为声学模型，大规模 N 元文法（N-gram）[3]作为语言模型，利用加权有限状态机（Weighted Finite State Transducer，WFST）[4]构造搜索空间进行解码操作。这些技术及其不断改善[5]导致了可以大规模商用的通用语音识别系统的出现。

语音合成从概念上是语音识别的逆过程，目标是将文本转换为语音。传统拼接式语音合成已经完全商用，可以合成具有高自然度和可懂度的语音。近年来，基于“源-滤波器”机理的参数化统计合成[6]受到越来越多的关注，这种方法在模型大小、个性化程度方面较拼接方法具有优势，近年来在声学模型训练[7]、基频建模[8]、深度学习[9]等方面产生一系列新技术，促进了它的成熟和广泛使用。

随着移动互联网时代的到来，人机对话系统的输入输出通道形式也变得越来越丰富。传统的键盘、鼠标、触摸板在需要精确输入的场景下仍然使用广泛，其它的输入通道，除语音之外，图像、手

[1] 知识图谱、语义网络等知识管理方式所处理的海量知识属于“开放知识”的管理，不在本文讨论范围之内。

势、触感、体感，乃至脑机接口，在更自然的移动应用中也受到了越来越多的重视。这些输入通道是连接物理信号与信息编码的纽带，有些具有文字编码特性，可以表达丰富的抽象语义，例如语音、光学字符识别（Optical Charactor Recognition）等；另一些则表达非文字性的语义信息，例如情绪、位置、行为类别等等。人与人对话中对信息的感知和交互往往是多模态的，因而人机对话系统中，将各类通道信息融合的多通道输入输出技术是被广泛接受的技术方向。例如 Ernst 和 Banks 采用统计优化的方法推动了触感与视觉感知的融合技术 [10]，微软开发的 Kinect [11]支持图像、手势、体感和语音识别的功能，实现了多通道信息的处理。多通道的优势是可以利用各种传感器和通道的互补性，使得人机交互更加灵活和便捷，提升输入输出的信息带宽，更符合人类交互的自然习惯。

抛开具体的通道模态，通道技术解决的问题是信号到编码的转换。虽然某些编码具有一定的语义作用，但一般情况下，编码并不等于最终的用户意图。而且在自然交互方式，尤其是多通道交互情况下，通道层的编码往往具有不确定性，整合协作和交互功效变成了一个瓶颈问题。这些都使得真实用户意图（语义）的解析很难通过提升单独通道的感知能力来彻底解决，而要涉及对话系统后端的融合、理解和控制，即控制层的认知技术。

认知型交互系统对于通道技术的一个重要观点是：经过通道获取的编码不是对用户意图的最终解释，而都应被视为用来推测用户最终语义或产生系统反馈的某种“特征”。在这种观点下，除信息融合之外，认知型对话系统要求对通道编码的不确定性进行显示的建模，以最大可能的传输信息，这是认知型对话系统框架下的通道技术与传统通道技术研究的不同点。

## 3 对话系统认知技术的范畴及其进展

如前文所指出，现代人机对话系统需要一个“控制层”将通道层的编码，与后端的知识连接起来，起到控制协调前后端的作用。“控制层”的主要功能包括：从输入编码中理解用户意图（语义理解）、交互逻辑的管理和控制（对话管理）以及意图向输出编码的转换（信息生成）。控制层在传统的机械式人机交互系统中是可以忽略的，因为用户的意图被输入手段精确的限定了。例如在 GUI 交互中，通过鼠标点击实现视窗的开闭是直接由输入层转换到内部知识管理。然而，自然人机对话系统要求机器也作为一个认知主体，可以按照人的方式和效率进行对话响应，这就使得控制层成为一个独立且不可或缺的模块，与这一模块相关的技术就是“认知技术”。认知技术的范畴主要涉及如下几类：

- 非精确条件下的理解

不确定性（或非精确性、不准确性），是人机对话通道的本质属性之一。语音识别本身由于噪声干扰、说话人语速口音等问题具有不可避免的错误。多通道输入的情况下，各个通道都有干扰产生不确定性。在 IO 层中的编码转换过程中的误差，再传递到语义解析层，就引发了语义解析的不确定性。另一方面，从认知角度，人类也自然的倾向于用非精确的信息进行交流，因为这会大大的增加信息传输的速度。在信息传输和语义本身具有不确定性的条件下，由机器对用户意图进行理解就成为认知技术的重要范畴之一。它与传统的“语义理解”或“自然语言处理”的根本不同就是将于不确定性纳入到研究范畴之内。

- 基于不确定性的推理及决策控制

人机对话系统的重要特征之一就是有效的多轮交互。在特定的任务下，基于系统运行状态和用户意图理解，尤其是不确定的意图理解，进行推理和对话决策，选择最有利于任务完成的反馈方式和反馈内容，是认知技术的另一重要范畴。交互策略的本质是人机对话的闭环控制技术，它会使得机器具有“推理和决策”这一认知主体特性。

- 交互自适应及进化

学习和适应能力是认知主体的另一重要特性。机器对用户在输入输出通道和控制层面的适应技术是认知技术的第三个范畴。在对话过程中，它既包括对于用户输入输出特征的底层自适应，又包括对于用户行为和交互逻辑习惯的自适应。另一方面，除了短期的自适应，认知控制技术还包括长期的系统“进化”，即从与人的长期互动中学习新的知识（如语义项）和行为方式的技术。

- 诱导式信息生成及传递

机器的对话决策是由反映机器意图的语义项所表示的，而语义信息的自然表达不仅要起到信息传递的作用，还要同时考虑认知引导层面的作用，能够引导用户更便捷、舒适的完成对话任务。将机器意图以符合人类认知习惯的方式自然的表达及传输出去，是认知技术的第四个范畴。

以下将对认知技术的四个范畴相关的概念、技术框架和进展进行详细介绍。

### 3.1 非精确条件下的理解

信号经过输入通道之后的表现形式是信息编码，系统需要进一步以此为特征去理解用户想要表达的意图。用户意图在研究中通常用“语义”这一概念来表达。作为对话系统语义概念的发展的重要一步，1999 年，Traum 发展了对话系统中的行为的概念，考虑了对话的轮次信息以及用行为来表达对话的意义，包括请求、询问等等行为 [12]。但是这种行为的表示和具体的语义信息是分离的。为了可以表达更具体的意思，一种简单有效的语义表达形式——对话语义动作（dialogue act) [13]被提出，它包括了一句话的行为以及其所带的若干简单的语义信息：

$$acttype(a = x, b = y, \cdots)$$

其中 $acttype$ 表示一句话的语义动作类型，$a = x$ 和 $b = y$ 表示的是语义动作涉及的属性和值，即 slot=value，被称为“属性值对”（slot-value pair）。同时，acttype 和这样的 slot-value pair 统一被称为语义项（semantic items）。

如 2.2 节所述，自然口语对话系统中的语音识别难以避免错误，且其规律性也很难发现。这就使得语音通道的输入具有非精确性。传统的优化观点认为，提升识别准确率，减低非精确性是实现有效语音理解的唯一途径。然而，从认知技术的角度去看，人类语言自身就具有高度的模糊性，认知科学的观点认为，允许使用模糊的表达手段可以避免不必要的认知负担，有利于提高交互活动的高效性和自然度。允许非精确输入，将使得信息的输入带宽大大提高，人机交互的自然性和高效性极大改观。因此，如何在非精确条件下实现有效的理解，即认知统计意图理解，是认知技术的重要研究范畴。

认知统计意图理解就是从非精确的编码输入中，得到准确的最优或多重用户意图理解。它和传统自然语言处理不同之处在于，可能存在多重通道的编码以准同步的方式输入，输入编码本身可能存在与用户意图无关的编码错误，且对应同一输入信号，通道层可能输出多种编码解释。多种编码解释是由于信息从输入通道中传输而产生的不确定性，这些不确定性与通道自身的性质或对话情境有关。保留合理的多重编码解释或利用多通道的非精确输入会为用户意图的理解和后续决策提供更多的信息，因而认知型统计意图理解范畴下，具有不确定性的输入通道的多重编码解释技术就成为重要的一环。

多重编码及置信度代表了输入不确定性，其表达形式可以有很多种。针对语音输入的一种简单有效的方法是采用 N 最佳列表（N-Best list），即对应于每句输入信号的 N 个最可能的编码序列及其对应概率。但这种表达方式不够紧凑，不同的编码序列之间往往有大量的重复信息，所以有人直接使用词混淆网络（word confusion network, WCN）来表达多重编码 [14]。图像信号输入特征的不确定性度量方法也各有不同。在基于图像分割的图像识别方法里面，分片的整体错误率（total error rate）和误分率（Misclassification error rate）分别可以作为一种不确定性度量方法 [15]。在图像根据实体分类的问题上，其不确定性度量可以是关于每类实体的概率 [16]，并且在图像理解的研究里，一种形式更加简单有效的不确定性度量方法是实体关于图像的后验概率 [17]。

早期的语义解析方法往往基于规则（rule based），例如商业对话系统 VoiceXML 和 Phoenix [18]，由于编码含有错误和概率分布，往往错误率很高。因而，数据驱动的统计方法，如隐矢量状态模型（Hidden Vector State）[19]、基于统计机器翻译的语义解析 [20]、条件随机场（Conditional Random Fields）[21]、深度学习[22]等等都取得了显著提高。从认知统计意图理解角度看，除对单一编码的解析之外，对于多重编码信息的语义解析是另一个新课题。支持向量机（Support Vector Machines）[23]就曾被用于含有多重编码的统计理解。当编码是紧凑的 WCN 网格形式的时候，基于网格的统计理解得到了比普通统计理解算法更准确的最优和多重理解结果[24]。

在认知统计意图理解的框架下，对于输入信息的质量衡量无法采用传统的单一确定编码的性能准则，如语音识别的字错误率或者语义解析的 F-score，它需要同时衡量多重编码的综合准确程度和置信度的可信水平。关于置信度的性能衡量，研究者通常采用归一化交叉熵（normalized cross entropy, NCE），这种准则可以衡量语音识别或语义解析的置信度的质量，但是不能有效衡量结果的正确率。语义项交叉熵（Item Cross Entropy, ICE），作为一种扩展方式，可以同时衡量准确率和置信度水平 [25]。

### 3.2 基于不确定性的推理和决策控制

推理和决策是对话系统中“控制层”的核心，又被称为“对话管理”。它决定着系统对用户的反馈 [~]，控制着整个对话的流向[26]。对话管理器有两个核心任务：一是维护系统的状态，称为对话模型；二是负责基于系统状态选择合适的动作回复给用户，称为动作选择。对话管理器可以分为两类：

- 基于规则的推理和决策

一个简单的对话管理方法是定义一组系统在对话中遵循的规则，在此框架下，系统一般会通过不断向用户提问，期望用户回答的方式来控制对话的流向。因此，这是一种系统主导（system-initiative）[27,28]的方式。在这种系统中，对话模型一般是有限状态机（finite state automation），有限状态机的每个结点代表系统的状态，与结点相连的弧代表用户可能的回答。这种系统的设计严重依赖于相关领域知识，从而导致系统的扩展十分困难。例如，向数据库添加数据或者删除数据都会引起有限状态机中结点和弧的增加和减少。其次，这种系统对语音识别和理解错误十分敏感，需要额外的组件来处理这些错误。

规则类推理的另一种方法是填表法（form-filling）[29]。在这种方法中，首先需要定义一组属性（slots），这些属性是用户在对话过程中可以提到的 ，然后将系统的状态分解到各个属性上。在对话的过程中，系统不断收集各个属性的值，然后更新系统的状态。这种方法给予了用户较大的主动性，具有较大的灵活性，但是对话管理器的设计者需要花费较大的时间和精力来设计对话策略，同时也不能较好地处理语音识别和语义理解错误。

典型的基于规则的对话系统有：为旅客提供航班信息的 ATIS 系统[30]，卡内基梅隆大学开发的 Let's Go!系统[31]等等。

- 基于统计的推理和决策

对话管理也可以看成是一个分类任务，即每个对话状态和一个合适的对话动作相对应。和其它有监督的学习任务一样，分类器可以从标注的语料库中训练得到。但是，在某状态下系统应该选择的动作不能仅仅是模仿在训练数据中同一状态对应的动作，而应该是选择合适的动作能够导致一个成功的对话。因此，把对话过程看成是一个决策过程更为合适，从而根据对话的整体成功来优化动作的选择过程[32]。因而这是一个规划问题，并且可以用强化学习[33]方法学习获得最优的结果。

在强化学习方法中，我们一般根据对话状态、系统的动作和回报来建模对话。假设当前的系统状态只依赖前一个系统状态以及用户的动作是完全可见的，则对话过程可以被看作一个马尔科夫决策过程（Markov Decision Process：MDP）[32]。但是在实际中，由于系统语音识别和语义理解都会产生错误，系统不可能确切知道用户的当前动作是什么，因此 Roy 等人提出用部分可观测马尔科夫决策过程（Partially Observable Markov Decision Process：POMDP）来建模对话过程[34,35]。

认知型人机对话系统的关键是认识到任何输入通道都具有本质上的非精确性。因此，认知型推理和决策不将输入视为确定的命令，而把它们视为一种观察到的特征，利用这些特征，系统可以推断用户的意图，并通过最大化某种收益值来量化的优化决策逻辑。POMDP 作为一种数学工具，为这样的推理和决策提供了重要的工程框架，它将反馈控制问题形式化，并使得数据驱动的交互逻辑优化成为可能。

#### 3.2.1 部分可观测马尔可夫决策过程

部分可观测马尔可夫决策过程是一个 8 元组 $(S,A,T,R,O,Z,\gamma,b_0)$。其中，$S$ 是机器的状态 $s$ 的集合，刻画了机器对用户意图和对话历史的所有可能理解；$A$ 是机器所有可能的动作 $a$ 的集合；$T$ 定义了一组状态转移的概率 $P(s_t \mid s_{t-1}, a_{t-1})$；$R$ 定义了一组瞬时收益函数 $r(s_t, a_t)$，表示特定时刻的特定状态下，机器采取特定动作的时候获得的收益；$O$ 表示所有可以观察到的特征集合，$Z$ 定义了基于状态和机器动作的特征转移概率，$P(o_t \mid s_t, s_{t+1})$；$0 \le \gamma \le 1$ 是强化学习的折扣系数；$b_0$ 是状态分布的初始值，又称为初始置信状态。

一个POMDP过程描述了机器和人交互进行决策的过程。一个典型的对话系统POMDP过程可由图 2 所示的动态贝叶斯网络（Dynamic Bayesian

[~] 这里的动作是指对话动作（dialogue act），即对话管理器的输入和输出都是对话动作。

Network）[26]：

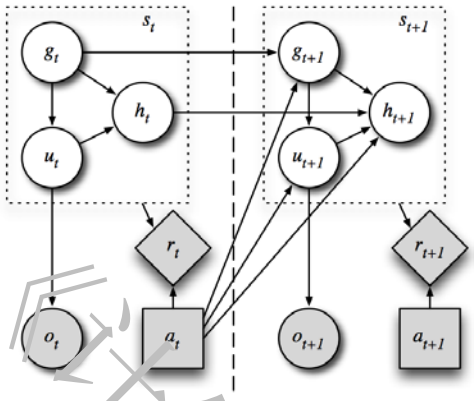


图 2　POMDP 示意图

在每个时间点$t$，POMDP 系统都处于某个未知的状态$s_t$。在自然人机对话系统中，这个“状态”必须能够描述三个方面的信息：用户的终极意图$g_t$，它代表了机器必须从用户那里获取的能够正确完成任务所需要的信息；最近的用户输入中包含的单句意图理解$u_t$，它代表了用户刚刚在时刻$t$说过的话，以及所有的对话历史$h_t$。这就使得实际的对话系统的状态可以分解为$s_t=(g_t,u_t,h_t)$，如图 2 中虚框部分所示，而这三部分在真实的人机交互过程中又都不是可以直接精确观测的[36]。系统在时刻$t$的全部状况则由所有状态的概率分布$b_t(s_t)=P(s_t)$表示，这个分布通常是一个离散分布，可简写为$b_t$，又被称为“置信状态”。需要强调的是，“置信状态”是分布而不是一个具体状态，它是对于系统全局状态的完整综合描述，包括了所有非精确的信息。基于置信状态$b_t$，机器会根据一定的策略选取机器行为$a_t$，基于此收获一个收益值$r_t$，并产生状态转移，形成新的状态空间。机器行为是用户可以观测的，而收益值取决于当前系统的状态$s_t=(g_t,u_t,h_t)$和机器行为$a_t$，一般是预先设计好或可以估计。新转移到的状态$s_{t+1}$也是不可见的，从统计上，它仅仅依赖于上一时刻的状态$s_t$和机器行为$a_t$。其中，观察特征$o_t$是通过识别和理解模块观察到的用户意图，表现形式是语义信息项。如 3.1 节所述，$o_t$会具有一定的不确定性，不同于真正的用户单句意图$u_t$，但从对话系统运行角度，它在统计上仅仅依赖于$u_t$。

由于 POMDP 提供了置信状态跟踪和策略优化的数学方法，它成为解决基于不确定性的推理和决策控制的重要工具。但人机对话的状态包括了大量的语义项和项值，用户意图、理解结果和对话历史的各种可能组合更使得状态空间的规模指数增长，一个不大的研究任务的状态空间都可能以百万计[35]，一般性的 POMDP 算法在理论和实践上都不可行。这使得 POMDP 在人机对话系统中有了更多新的需要解决的问题，构成了认知技术的重要部分。

3.2.2　对话模型：推理及置信状态跟踪

置信状态的跟踪本质上是对概率分布$b_{t+1}$的估计，大多数研究都是基于贝叶斯公式的展开进行分项研究。由图 2 中给出的统计相关关系假设，可以得出置信状态进行统计跟踪的基本公式如下：

$$\begin{aligned} b_{t+1}(g_{t+1},u_{t+1},h_{t+1}) = \\ \eta P(o_{t+1}\mid u_{t+1})P(u_{t+1}\mid g_{t+1},a_t) \\ \sum_{g_t} P(g_{t+1}\mid g_t,a_t) \\ \sum_{h_t} P(h_{t+1}\mid g_{t+1},u_{t+1},h_t,a_t)b_t(g_t,u_t,h_t) \end{aligned} \quad (1)$$

上面公式中展开的各项代表不同的物理含义，对应于人机对话的不同层次的模型，包括：

1. 观察模型$P(o_{t+1}\mid u_{t+1})$：对语音识别和语义理解中可能的误差进行了建模。
2. 用户模型$P(u_{t+1}\mid g_{t+1},a_t)$：表现了在一定的用户意图和系统反馈下，用户可能表达的具体语义，这是对用户的行为特征的建模。
3. 意图转换模型$P(g_{t+1}\mid g_t,a_t)$：表达了用户的意图在对话过程中转换的概率。
4. 历史模型$P(h_{t+1}\mid g_{t+1},u_{t+1},h_t,a_t)$：表达了系统对对话状态历史的记忆。

以上的模型每一项都有较高的复杂度，为了使得 POMDP 模型能够可计算，近似算法就成为置信状态跟踪的核心之一。近年来，两类近似算法被广泛使用：

- N-best 近似

  原始的置信状态$b_t$描述的是所有可能的状态

⎵ 动态贝叶斯网络的每个节点表示一个随机变量，节点之间的箭头表示随机变量之间的统计相关性，A→B 表示 B 依赖于 A，没有箭头的节点之间是条件独立的。阴影节点代表可观测的值，空白节点表示隐变量。

$s_t$ 的概率分布，N-Best 近似的基本原理是用一些最可能的状态列表来代替近似整个状态空间。这意味着只有那些有较高概率的对话状态才会被有效描述，而其它的状态只有很小的概率。这种框架下的一个典型例子是“隐信息状态”模型[35]，它将相似的用户意图聚类，形成基于树结构的聚类分割，这些类可以随着对话过程的继续进行动态的分割和合并。而置信状态的跟踪仅仅在类的级别进行，这就大大减小了计算复杂度。在类似的框架下，如何有效的进行聚类[37,38]，状态空间剪枝算法[38,39]，状态的概率形式[40,41,42]等等问题都得到了较多的研究，使得 N-Best 近似能够成功应用于小规模的真实世界对话任务[35]。N-Best 近似的框架从原理上可以看做是有 N 个基于精确输入的对话管理器在并行运行，不同的对话管理器对应于对用户所说的内容的不同理解。

- 因素分解近似

与 N-Best 近似算法不同，因素分解是在结构上对用户意图进行进一步的统计分解。例如对于一个旅游信息系统，用户的意图可能是餐馆或旅店，而这两种不同类型的意图又分别对应于不同的具体语义项，例如菜品、星级等等，于是从先验知识出发，可以假定餐馆相关的用户意图与旅店相关的意图是完全独立，即可以分别独立跟踪。由于这种分解，用户意图可以在语义项的层次进行分别跟踪，这使得完整的概率分布跟踪成为可能，在独立性假设合理的情况下，会优于 N-Best 的近似方法[43,44]。由于因素分解近似采用完整的概率分布，机器学习领域很多标准的置信跟踪算法都可以应用[45]。一些改进算法在近年也被提出来，使得因素分解近似也可以对一定的意图相关性进行建模[43,46]。因素分解近似也已经成功的应用于小规模的真实世界对话任务[47]。

以上的状态跟踪算法都是基于公式(1)进行的。状态跟踪本身也可以看做是一个分类加置信度的机器学习问题，引发了很大的研究兴趣[48]。从认知技术角度看，置信状态是人机对话控制层的输入信号。以上算法的共性都是将输入通道产生的不确定性予以保持，并通过跟踪过程借助对话运行中产生的上下文信息来逐步消除或降低不确定性。

3.2.3 动作选择：基于强化学习的策略优化

在 POMDP 框架下，对话系统的“策略”是一个从置信状态 $b$ 到机器行为 $a$ 的映射。这个映射既可以是确定性的映射函数[35]，表示为 $a=\pi(b)$，也可以是随机映射[49]，表示为给定置信状态产生特定机器行为的概率，$\pi(a|b)\in[0,1]$ 且 $\sum_a \pi(a|b)=1$。除这两种主流的状态表示之外，也有一些其它形式的策略表示，如有限状态控制[50]，观察序列到机器行为的直接映射[51]等。

基于数据对策略进行统计学习是认知技术的重要体现，强化学习[33]就提供了这样的框架。其基本原理是对每个对话时刻（轮回）定义收益值，之后找到合理的策略（映射），使得统计上整个会话完成之后的加权收益最大。数学上，从置信状态 $b_t$ 出发，按照确定性的策略 $\pi$ 进行对话的情况下，加权收益值的期望可以表达为一个递归函数，最优的策略 $\pi$ 就是满足加权收益期望最大的策略。尽管精确[52]和近似[53]的 POMDP 策略优化算法都已经在传统强化学习的文献中被提出，这些标准算法都无法在真实世界对话系统的尺度上运行。这是由于用户的意图、可能的机器行为和用户输入的组合过于庞大，即使是一个中等规模的系统，组合数可以很轻易的达到 $10^{10}$ 以上[26]，这就使得针对认知主体的强化学习与传统强化学习有根本的不同，必须采用新型的近似算法才能得到实用系统。

近似算法的基本思路是假定状态空间中相邻的点可以对应同样的机器行为 $a$，这就需要将整个状态空间进行分割，每个分块中的所有点对应同样的最优机器行为。尽管进行了分割，精确的 POMDP 策略在真实系统中仍然由于状态空间规模过大是不可计算的。考虑到在真实对话系统中，虽然可能性众多，但实际只会有很小部分的置信空间和机器行为会被用到，如果在这个较小的子空间中进行计算，POMDP 的策略优化就变得可行了。这就引入了所谓“摘要空间”的概念[54]。在这一框架下，对话系统运行的时候，置信状态的跟踪在主状态空间进行，在状态转移完成后，主空间的置信状态 $b$ 被映射到摘要空间的置信状态 $\hat{b}$ 和摘要机器行为集合 $\hat{a}$，之后就通过策略函数选择 $\hat{b}\rightarrow\hat{a}$ 的最优映射，之后，利用一些启发性的知识再将摘要机器行为 $\hat{a}$ 映射回正常的机器行为 $a$。这样，策略的优化和决策确定都是在摘要空间完成。

摘要空间技术的一个核心问题是如何将摘要机器行为映射到主空间，得到完整的机器行为。一个

简单方法是采用对话行为的类型作为摘要机器行为，而到主空间行为的映射仅仅自动的将此对话行为类型与具有最高的置信度的语义项结合[35]。这种方法的好处是可以全部自动化，不足之处是可能出现逻辑错误。另一类方法是建立人工规则[55]或马尔可夫逻辑网络[56]，这类方法可以将先验知识有效的引入而且在训练最优策略的过程中可以加快收敛速度，但人工规则会有正确性风险，可能把最优的机器行为遗失。

摘要空间技术的第二个核心问题是如何抽取状态和机器行为的特征供计算使用。对机器行为而言，可以简单的用二值特征来表示某个对话类型或语义项是否出现，一般情况下会有 20-30 维的特征，每一维度表示一个独立的摘要机器行为。对于状态而言，特征往往具有不同的数据类型，包括实值特征、二值特征或类别特征等等，具体的特征物理含义包括用户意图的 N-Best 猜测、数据库匹配的条目数、对话历史等等。状态特征不一定仅仅限于置信状态的特征，它也包括一些外部特征，如数据库的信息等。

给定摘要空间后，对话策略就可以表示为确定性的映射 $\pi(\hat{b}) \to \hat{a}$ 或者随机映射 $\pi(\hat{b},\hat{a}) = p(\hat{a}\,|\,\hat{b})$，在随机映射情况下，最终的机器行为是从条件概率中采样得到。这些映射函数的学习是策略优化的核心内容，占主流的方法都是通过优化 Q 函数发现最优的策略映射。很多丰富的优化方法被提出来，例如非确定规划[53,57]、蒙特卡罗优化[58]、自然梯度优化[49]等等。

3.2.4 用户模拟器

如上节所述，对话策略是系统对话状态到对话动作的一个映射，而一个好的对话策略需要系统在反复的交互中学习得到。理论上训练统计对话系统可以使用真实的用户或者使用系统-用户交互的语料，但是对于现实的大规模应用领域来说，对话的状态空间是十分巨大的，使用上述两种方法需要太多的人力或者超大规模的训练语料。因此，建造一个用户模拟器 [59]用以代替人和对话管理器交互是十分必要的。有了机器模拟的用户，就可以进行海量的多轮交互的完整对话，使得统计对话管理器的学习或评估成为可能。其基本思想是：以机器（用户模拟器）代替人来训练机器（对话管理器），用以得到最终可以与人进行自然交互的机器（对话管理器）。虽然这种思想受到了一定质疑，但从统计对话系统的研发角度看，用户模拟器与对话管理器往往采用不同的模型进行独立的训练，用户模拟器可以比静态语料更充分的遍历可能的状态空间，对统计对话管理器的训练，尤其是从无到有的初始化具有重要的作用。

用户模拟器本质上是一个可以和对话系统直接交互的用户决策系统，是对话管理器的逆过程，代表了真人用户在交互过程中的响应。它既可以由规则确定，也可以引入数据驱动的方式从语料库中学习得到。图 3 显示了对话管理器和用户模拟器的交互过程。

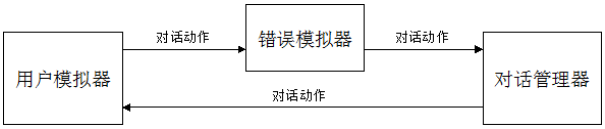


**图 3 用户模拟器对话管理器的交互过程**

如图 3 所示，在使用用户模拟器训练或者评估对话管理器时，每一轮对话 [1]中，用户模拟器的输出经过错误模拟器后传递给对话管理器，然后对话管理器根据其策略选择一个动作回复给用户模拟器。错误模拟器是一个模拟语音识别和语义解析错误的模型。给定了用户模拟器和错误模拟器之后，就可以通过生成海量的对话来训练对话管理器的参数，或对确定参数的对话管理器的性能进行评估。

根据对话建模的抽象层级不同，用户模型可以分为以下几类：语音层级 [60]、文本层级 [61]及语义层级 [62]。在最近几年，研究人员更加关注于语义层级用户模型的研究，图 3 就显示了这一层级的用户模拟器交互过程，即各模块的接口是“对话动作”。以早期的 N-grams 模型 [63]用户模拟器模型为例，它假设在 $t$ 时刻用户模拟器的动作 $u_t$ 仅仅与系统的对话历史和之前的用户模拟器动作有关，且采用 bigram 来近似完整的对话历史：

$$u_t = \arg\max_{u_t} p(u_t \,|\, a_{t-1}, u_{t-1}, a_{t-2}, u_{t-2}, \ldots, a_1, u_1)$$
$$\approx \arg\max_{u_t} p(u_t \,|\, a_{t-1}, u_{t-1})$$

其中，$a_t$ 表示 $t$ 时刻对话管理器的对话动作。这个模型完全是基于概率的，并且与具体领域无关。但是，它没有对用户模拟器的动作 $u_t$ 作任何限制，任何用户动作都是系统当前动作的合法回复，因而导

[1] 一轮对话是指用户（或者用户模拟器）和系统各说了一句话。

致用户目标经常改变或者经常重复之前的动作，这样产生的对话一般都比较长。为克服早期模型的问题，研究者们提出了一系列新方法，如基于连接图的模型 [64,65]、贝叶斯网络 [66]等等。近年来，基于议程（agenda）的模型[62]及其扩展，如逆强化学习（Inverse Reinforcement Learning, IRL）[67]等是受到较多使用的方法。

用户模拟器的性能的好坏能直接影响到对话系统的性能的分析和所学策略的好坏。用户模拟器的性能评估还是一个开放问题，目前还没有一致的衡量指标[68]。Pietquin和Hastie[69]提出了一个好的用户模拟器的性能评价指标需要满足的若干条件。在过去的十几年中，有许多不同的评价指标被提出。精度和召回率被用来衡量用户模拟器对用户动作的预测能力[70]，KL 距离衡量真实对话过程和用户模拟器对话过程中对话动作分布的距离[66,68]，任务完成度、对话时间和平均长度等指标可以用来衡量用户模拟器的整体性能[66][68]等等。Pietquin等人[69]对用户模拟器评价进行了综述，并提出性能评价方法可分为局部指标（即描述单轮的统计特性）和全局指标（即衡量整个对话层级的统计特性）。

### 3.3 交互自适应及进化

对环境和交互过程的学习和适应是认知主体的智能的重要体现，前节所述的对话策略也可以从广义上看做是在交互过程中机器对用户反馈的自适应回答，本节主要关注其它的自适应技术。

以口语对话系统为例，对于语音输入通道的自适应是指在有少量用户数据的情况下，改变系统参数或算法，提升对特定用户的语音识别性能。这包括了对声学环境，包括噪音、口音、用户说话方式等的自适应和对语言领域的自适应。声学自适应是语音识别领域发展较长时间的技术，包括对特征估计一定的环境相关的回归变换，如 CMLLR，以及对模型参数进行自适应变换，如 MAP、MLLR 等技术，作为自适应的扩展，一系列自适应训练技术，如 CAT、SAT 等，也被提出，用于改善基础声学模型的质量，相关的内容可参见综述[71]。语言模型的自适应的主要目的是使语音识别的搜索空间更偏向于用户所谈及的领域和词表。在基于语法的识别系统中，这种自适应往往是通过对话状态选择一系列特定的语法进行识别，优点是简单易行，在选择正确的情况下识别率高，缺点是易出错，一旦出错错误率较高，且难以恢复。基于语言模型的识别系统往往会得到比较稳定的结果，常见的自适应方法包括模型组分差值自适应、动态缓存模型、MAP 自适应等方法，相关的内容可参见综述[72]。但传统的声学和语言模型自适应都是离线进行的，在人机对话系统中进行在线的数据收集和自适应则是从交互自适应角度出发提出的新课题，这其中主要涉及渐进性的模型自适应和跨模块联合自适应等问题，相关的研究有所展开，但尚比较零星，且在对话系统框架下进行研究的还较少。

除输入之外，人在对话过程中也会对输出进行一定程度的自适应，如强调信息、改变语速和风格等等，以使得对话交互更为自然。对话系统的输出通道自适应主要是指对语音合成的风格自适应。由于基于 HMM 的统计语音合成的使用，这种自适应变得可行，对于说话风格的自适应方法大都是语音识别自适应的扩展，如 CMLLR，MAPLLR 等，相关内容可参加综述[73]，近年来，对于丰富上下文建模也引起了研究界的重视，产生了一些新方法，如上下文自适应[74]等，较好的实现了对特定信息的强调。

以上“自适应”的概念一般是指在一个对话内部的适应，除这种短期的即时适应外，长期的适应，也即“进化”，也是认知能力的重要体现。这主要体现在系统能从长期的互动中学习新的知识，这是研究的前沿，2012 年起设立的欧盟第 7 框架的 PARLANCE 项目，要求对话系统能根据交互总结并学习新的语义项，并把它用于未来的交互过程。在这一计划支持下，对话策略的进化算法有了较大的发展。例如[75]就采取了随机的初始对话策略，通过在线的用户交互不断自动更新对话策略，最终达到了较好的策略。这一研究开启了大数据情况下，采用在线方式训练交互逻辑的先河。

### 3.4 诱导式信息生成及传递

对话管理模块的输出是表示语义信息的对话行为，将其转化为输出通道信号的过程就是信息生成，主要包括自然语言生成、语音合成及其它通道信号合成。一个具有认知能力的系统应当在语义表现上也具有认知主体的诱导、自然、舒适等特点，因而信息生成的研究也是认知技术的研究范畴。

自然文本的生成需要解决“说什么”、“如何说”的问题，由于对话管理在系统级别负责“说什么”的问题，“如何说”就成了自然文本生成的核心问题。最简单的生成方式可以是预制文本（Canned Text）

或模板填充，这些算法往往只是基于规则来决定产生何种字符，所有功能基于字符串层，不涉及句法及文本计划层面[76]。更符合认知特点的方法仍然是基于语言学的方法，例如Penman[77]和KPML[78]用大量相互关联的语法混合词汇、语法、语义，使用明确的语言理论——系统功能语言学，利用语言的实用功能，说话人的态度，和实际观众的关系来生成文本。除基于规则之外，近年来，基于统计的方法也得到了重视，例如Nitrogen用过度生成和剪枝策略，由弱短语结构语法生成数以百万计的满足输入要求的可能的句子[79]。文本生成领域的更多详细内容可以参见[80]。

机器生成信息的自然、舒适性的本质是拟人化的情感表达，可以归为情感计算的领域，它代表了语义信息外的“言外之意”。一般借助文本[81]、语音[82]、图像（表情）等方式综合表达，详细的内容可参见[83]。

## 4 认知型人机对话系统未来研究方向

随着移动互联网的发展，传统的机械式对话系统已经不能满足人类自然交互的要求，大规模的自然人机交互需求被极大的激发出来。新型的人机对话系统既要实现基于非精确的自然语言的交互，又要能有效的完成用户任务，而不仅仅是随意的聊天娱乐交互。这就要求机器具有“认知主体”的特性和完成有目的交互的能力，因而相关的认知技术的研究也正在成为人机交互研究的一个新领域。认知技术的研究虽然已经展开，尤其在近年有较大的发展，但在算法研究、工程实践、范畴框架等方面还面临许多挑战。

### 4.1 POMDP模型的实用性

POMDP 是认知技术中实现基于不确定性的推理和决策控制的重要理论基石，它通过对状态不确定性进行明确的概率建模和基于强化学习的统计策略优化方法使得对话管理可以在数据驱动的框架下进行研究。如前所述，POMDP 过程极为复杂，实际使用需要有各种近似以使其实用。虽然这一方向已得到很大发展，但众多挑战还没有根本解决：

- 真实世界状态空间的定义

真实世界的任务型对话虽然目标明确，但在自然对话情况下，状态空间并不仅仅取决于先验的任务设定（如数据库槽值），还可能受到用户个性、数据库查询结果等的影响，确定合理和全面的状态空间是 POMDP 能够实际使用的前提。这一挑战本质上需要在真实任务上，面对真实用户实现对话系统。

- 状态复杂度与置信状态跟踪的折衷

POMDP 的复杂度根本上取决于状态空间的大小，而真实世界任务中的状态千变万化，如何采用合适的 POMDP 结构和算法是其能否实用的重要因素。状态空间太小无法描述真实世界的复杂任务，状态空间太大则使得置信状态跟踪不可计算，这是 POMDP 应用的核心问题。针对这一问题，对状态空间的结构化描述成为重要的思想。目前的主要方式还是依据先验知识建立图模型来表示状态直接的相关性，还没有有效引入数据驱动的统计学习方法。而机器学习领域，图模型的结构学习（structured learning）已经得到长足发展，未来研究中，基于数据驱动的结构学习将可能为 POMDP 状态空间压缩提供有力的方法论支持。另一方面，将置信状态的跟踪独立抽象为有监督学习的问题是一个重要的思路，在这一思路下，置信状态跟踪不再受限于贝叶斯公式，也使得基于大规模离线数据的学习更为方便，这将有效的促进置信状态跟踪算法的发展。

- 在线学习与进化

POMDP 中，策略的统计训练往往需要数千或数万的对话样本，这使得策略的学习速度缓慢，只能离线进行，一旦训练好后就无法改变。然而，认知能力的一个重要表现就是可以在线的进化。因此，实现在线的学习和进化，是个重要的研究方向，而这就需要系统策略的学习可以用渐进式的方式快速进行。POMDP 策略的快速在线训练需要涉及两个方面的研究：一是在线更新算法，这主要是解决通过小样本进行鲁棒的策略训练，与机器学习中的小样本学习不同，这里的小样本学习需要在强化学习而非监督学习的框架下展开。一些随机过程模型，如高斯过程，已经被用来研究这个问题[75]，但尚没有达到理想的效果。第二方面是需要有即时更新的回报函数，这将在下节讨论。

POMDP 的进化不仅体现在策略优化上，本体进化是更实际的一个方面。认知型人机对话系统的运行是基于特定的本体和语义项定义系统的，目前的系统设计的前提都是预先定义完备的本体、语义项和槽值数据库，即静态本体。然而，真实的人机

 DiMarco Chrysanne. Natural language generation. https://cs.uwaterloo.ca/~jchampai/CohenClass.en.pdf

交互过程中，往往会出现数据库的新项目，或要求发现新的语义项和本体关系等，这些基础结构的变化会很大的影响语义理解和对话决策。如何从与用户的交互中，学习这些“新知识”并把它有效的融合到现有的模型中，是“学习能力”的另一个重要体现。对于这一挑战，一种较可行的框架是将本体学习的问题归入知识管理模块，借用知识图谱中的本体学习方法，而在这一过程中引入交互中上下文的信息记录作为辅助特征促进本体的学习。

### 4.2 对话系统的评估和回报函数表示

从模块级别看，对话系统中 IO 层的技术（识别/合成）评估已有较成熟的方法和指标（如文字错误率、语义解析准确率、自然度主观评估等），而基于 POMDP 的认知算法，由于其输入和输出都无法观测而很难进行显示的主客观评估，往往只能通过对话系统的整体性能评估来表现其性能。作为双向信息连续交互的系统，这些评价最终是在特定任务下对一系列决策及其结果的综合评估。目前整体系统的评估主要有三种形式：用户模拟器、假想用户、真实用户。用户模拟器可以高效生成大量样本，因而被广泛使用[35]，但其交互方式与真实用户仍然可能有很大差距。假想用户是指实验室人员或受雇人员按照一定的任务指导进行测试，这些用户是真人，因而会更好的反应对话系统的性能，例如亚马逊的众包系统 Amazon Mechanical Turk 就被广泛使用[84]。但由于这些用户的真正目的并非完成任务，在其行为会与真实的使用情况有很大差距，他们对系统的主观评估也会有较大偏差，这在没有监督机制的情况下更为明显[75]。最有效的方式是真实用户的评测，如 CMU 的公交系统评测[47]，但任务仍然比较简单，不够实用。未来认知型对话系统的发展必然需要更多真实世界的交互测试平台。

POMDP 的一个核心优势就是通过强化学习实现数据驱动的策略学习，而这使得回报函数的设计在这一框架下具有举足轻重的地位。对话系统的评估指标不仅为了我们了解性能，还应和回报函数高度关联才能使统计训练的策略真正有效。理论上，最有效的指标当然是最终用户满意度，但这种主观指标受任务、用户类型、对话流程等多方面影响，且很难大量获得，无法用于训练，因而客观评价指标的研究和使用是未来重要的课题。一种典型的研究思路是采用一系列客观可测的特征去拟合最终的用户满意度，这被称为“PARADISE”框架[85]。目前，对话成功率和对话轮数是广泛使用的客观评价特征指标，但被认为过于粗糙，需要在回归框架下采用更精细准确和易用的客观评估特征，尤其是每个轮次上的回报函数定义具有很大的实际意义。对此挑战，一种较可行的思路是采用非语义的其它信息作为辅助特征构造回报函数，例如对声音和文本的情绪分析就是一种典型补充特征。更一般的，在移动互联应用下的对话系统，充分利用各种多模态反馈信息将是回报函数构造的重要途径。

### 4.3 对话系统的认知自然度

“自然交互”是认知型对话系统的目标，也是认知技术合理使用的结果。除了前述的自然语义交互方面的内容之外，认知型对话系统对其它方面的自然度要求也成为新的挑战。

情绪的检测和传达是人类自然交互中不可或缺的一环。情感计算已经发展多年，在情感分类和特征建模[86,87]、情感识别[88,89]、情感表达[90]等方面都有了较大进展，将情感分析应用于对话系统也得到了产业界的重视，呼叫中心中利用情感分析来获取用户满意度也得到了应用。未来研究中，如何将情绪等非语义的自然交互信息全面引入对话系统，是一个重要方向。这其中需要具体解决的重要问题包括：面向交互任务的情感、情境的量化分析感知及信息融合；情感的表达模型和高表现力的情感合成（主要是语音和图像）；引入情感因素的对话状态空间定义及强化学习算法等。情感计算中的情感维度空间模型[91]等提供了将情感特征量化的良好手段，将利于将情感信息引入 POMDP 框架。而在多模态交互条件下，结合文本、视频、音频进行联合的情感分析将是未来实现情感计算与对话系统结合的重要方法。

目前在国内外所有对话系统研究中，都有一个最基本的假设：人机交互的一个轮回必须是一个“句子”。但这种一问一答的方式与实际人类自由对话的方式相去甚远。很多心理学文献都明确指出人类的交互是渐进性的[92]。而且以整句为处理单位会使整个人机对话的时间变长，变得不自然，也会影响用户，使他们对目标的注意力下降。更重要的是，目前的研究全部把轮回检测看成与对话管理无关的独立任务，并没有尝试去研究轮回检测对整体对话系统的性能影响。尤其在较大尺度和真实环境的对话系统中，往往不是由于机器听不懂用户或者不知道如何反馈导致了对话失败，而失败往往来源于机器不知道何时对用户反馈或者用户不知道什么时候该对机器说话。这个现象意味着目前的对

话系统研究只关注“反馈什么”，却缺失了另一个重要的交互研究课题：“何时反馈”，因此，基于自然轮回的对话系统是认知技术未来研究的另一个重要课题。对此，一种可行的思路应是在对话管理器之外，引入新的“轮回管理器”，独立的对轮回和时序问题进行研究。

### 4.4 大规模真实世界认知型对话系统

虽然认知型人机对话系统已经在若干真实世界系统中得到过实现，但其系统都还是小规模或非真实的系统。例如CMU的Spoken Dialogue Challenge中的公交信息查询系统[47]虽然是在匹兹堡市运行的真实系统，但其处理的任务仅仅是公交车的站点和时间信息，任务规模很小；而欧盟CLASSiC项目[48]中的餐馆查询系统，虽然涉及的数据信息较多任务规模较大，但相关系统的运行还是基于招募的测试者而非真实的用户，这也使得对话策略的训练和对话系统的评估都受到了影响。认知技术至今仍然还没有在大数据真实条件下得到完整的实践验证。因此，在已有研究的基础上，面向真实世界的大规模任务搭建并运行完整的认知型对话系统，并与传统的机械式对话系统对比是认知技术的实验方向，也是应对前述各种挑战的必要的实测平台。

另一方面，大规模真实世界的认知型对话系统还对理论研究具有重要的推动作用。从深度理解、对话决策和自适应三个范畴来说，真实世界的大数据和实时运行都会带来更多的理论问题。例如多通道联合作用下的非精确信息理解，大规模状态空间下的 POMDP 模型，用户模型的精确学习，跨系统模块的交叉自适应（例如根据对话状态对语音识别模块进行在线自适应），实时在线的自适应和个性化等等。这些理论问题的产生都是由于真实世界的人机交互系统的最终性能是各个模块的有机叠加，而非某个模块的单独性能。因此，设计并实现大规模真实世界认知型对话系统也将有力促进从系统层面进行理论研究。对真实世界系统的需求使得与工业界的结合在认知型对话系统的研究中具有尤其重要的意义。

## 参考文献

[1] Dong Shi-Hai, Wang Heng. Human-computer interaction. Beijing: Peking University Press. 2003.(in Chinese)

董士海, 王横. 人机交互. 北京: 北京大学出版社, 2003.

[2] Dahland George E, Yu Dong, Deng Li, and Acero Alex. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech & Language Processing, 2012, 20(1): 30-42.

[3] Federico M., Bertoldi N., and Cettolo M. Irstlm: an open source toolkit for handling large scale language models. //Proceedings of the Annual Conference of the International Speech Communication Association(InterSpeech). Brisbane, Australia, 2008: 1618-1621.

[4] Mohri M., Pereira F., and Riley M. Weighted finite-state transducers in speech recognition. Computer Speech & Language, 2002, 16(1): 69-88.

[5] Senior Andrew and Lei Xin. Fine context, low-rank, softplus deep neural networks for mobile speech recognition. //Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP). Florence, Italy, 2014.

[6] Zen Heiga, Tokuda Keiichi, and Black Alan W. Statistical parametric speech synthesis. Speech Communication, 2009, 51(11): 1039-1064.

[7] Wu Y. J. and Wang R. H. Minimum generation error training for hmm-based speech synthesis. //Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP). Toulouse, France, 2006.

[8] Yu K. and Young S. Continuous F0 modelling for HMM based statistical speech synthesis. IEEE Transactions on Audio, Speech and Language Processing, 2011, 19(5): 1071-1079.

[9] Zen H., Senior A., and Schuster M. Statistical parametric speech synthesis using deep neural networks. //Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP). Vancouver, Canada, 2013.

[10] Ernst Marc O and Banks Martin S. Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 2002, 415(6870): 429-433.

[11] Zhang Zheng-You. Microsoft kinect sensor and its effect. Multimedia, IEEE, 2012, 19(2): 4-10.

[12] Traum David R. Speech acts for dialogue agents. Foundations of rational agency, Berlin, Germany: Springer, 1999: 169-201

[13] Thomson Blaise. Statistical methods for spoken dialogue management. Berlin, Germany: Springer, 2013.

[14] Hakkani-Tür Dilek, Béchet Frédéric, Riccardi Giuseppe, and Tur Gokhan. Beyond asr 1-best: Using word confusion networks in spoken language understanding. Computer Speech & Language, 2006, 20(4): 495-514.

[15] Wu Jin-Chu, Halter Michael, Kacker Raghu N, Elliott John T, and Plant Anne L. Measurement uncertainty in cell image segmentation data analysis. Maryland: NISTIR 7954, National Institute of Standards and Technology,

⁔ http://www.classic-project.org/

2013.

[16] Haralick Robert M, Shanmugam Karthikeyan, and Dinstein Its' Hak. Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on, 1973, (6): 610-621.

[17] Shotton Jamie, Winn John, Rother Carsten, and Criminisi Antonio. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision, 2009, 81(1): 2-23.

[18] Ward Wayne. Understanding spontaneous speech. //Proceedings of the workshop on Speech and Natural Language, Cape Cod, USA, 1989: 137-141.

[19] He Yulan and Young Steve. Semantic processing using the hidden vector state model. Computer speech & language, 2005, 19(1): 85-106.

[20] Wong Yuk Wah and Mooney Raymond J. Learning synchronous grammars for semantic parsing with lambda calculus. //Annual Meeting-Association for computational Linguistics, 2007, 45(1): 960-967.

[21] Wang Ye-Yi and Acero Alex. Discriminative models for spoken language understanding. //Proceedings of the Annual Conference of the International Speech Communication Association(InterSpeech). Pittsburgh, USA, 2006:1766-1769.

[22] Yao Kai-Sheng, Zweig Geoffrey, Hwang Mei-Yuh, Shi Yang-Yang, and Yu Dong. Recurrent neural networks for language understanding. Lyon, France, 2013.

[23] Mairesse Fran?ois, Gasic Milica, Jurcícek Filip, Keizer Simon, Thomson Blaise, Yu Kai, and Young Steve. Spoken language understanding from unaligned data using discriminative classification models. //Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP). Taipei, Taiwan, China,2009: 4749-4752.

[24] Henderson Matthew, Gasic Milica, Thomson Blaise, Tsiakoulis Pirros, Yu Kai, and Young Steve. Discriminative spoken language understanding using word confusion networks. //Proceedings of the IEEE Spoken Language Technology Workshop(SLT). Miami, USA, 2012: 176-181.

[25] Thomson Blaise, Yu Kai, Gasic Milica, Keizer Simon, Mairesse Francois, Schatzmann Jost, and Young Steve. Evaluating semantic-level confidence scores with multiple hypotheses. //Proceedings of the Annual Conference of the International Speech Communication Association(InterSpeech). Brisbane, Australia, 2008: 1153-1156.

[26] Young Steve, Gasic M, Thomson Blaise, and Williams Jason D. Pomdp-based statistical spoken dialog systems: A review. Proceedings of the IEEE, 2013, 101(5): 1160-1179.

[27] Walker Marilyn and Whittaker Steve. Mixed initiative in dialogue: An investigation into discourse segmentation. //Proceedings of the 28th annual meeting on Association for Computational Linguistics, Pittsburgh, USA, 1990: 70-78.

[28] Chu-Carroll Jennifer and Brown Michael K. Tracking initiative in collaborative dialogue interactions. //Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Madrid, Spain, 1997: 262-270.

[29] Zue Victor W and Glass James R. Conversational interfaces: Advances and challenges. Proceedings of the IEEE, 2000, 88(8): 1166-1180.

[30] Stallard David and Bobrow Robert. Fragment processing in the delphi system. //Proceedings of the workshop on Speech and Natural Language, New York, USA, 1992: 305-310.

[31] Raux Antoine, Langner Brian, Bohus Dan, Black Alan W, and Eskenazi Maxine. Let's go public! taking a spoken dialog system to the real world. //Proceedings of the Annual Conference of the International Speech Communication Association(InterSpeech). Lisbon, Portugal, 2005: 885-888.

[32] Levin Esther, Pieraccini Roberto, and Eckert Wieland. Learning dialogue strategies within the markov decision process framework. // Proceedings IEEE Workshop on Automatic Speech Recognition and Understanding,Santa Barbara, USA, 1997: 72-79.

[33] Sutton Richard S and Barto Andrew G. Reinforcement learning: An introduction. Cambridge, UK: Cambridge Univ Press, 1998.

[34] Roy Nicholas, Pineau Joelle, and Thrun Sebastian. Spoken dialogue management using probabilistic reasoning. //Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, China, 2000: 93-100.

[35] Young Steve, Gasic Milica, Keizer Simon, Mairesse Fran?ois, Schatzmann Jost, Thomson Blaise, and Yu Kai. The hidden information state model: a practical framework for pomdp-based spoken dialogue management. Computer Speech and Language, 2010, 24(2): 150-174.

[36] Williams Jason D and Young Steve. Partially observable Markov decision process for spoken dialog systems. Computer Speech and Language, 2007, 21(2): 393-422.

[37] Kim K., Lee C., Jung S., and Lee G. A frame-based probabilistic framework for spoken dialog management using dialog examples. //Proceedings of Special Interest Group on Discourse and Dialogue(SigDial). Columbus, USA, 2008: 120-127.

[38] Gasic M. and Young Steve. Effective handling of dialogue state in the hidden information state POMDP dialogue manager. ACM Transactions on Speech and Language Processing, 2011, 7(3):4.

[39] Williams J. Incremental partition recombination for efficient tracking of multiple dialogue states. //Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP). Dallas, USA, 2010: 5382-5385.

[40] Higashinaka R., Nakano M., and Aikawa K. Corpus-based discourse understanding in spoken dialogue systems. //Proceedings of the Annual Meeting of the Association for Computational Linguistics(ACL). Sapporo,

Japan, 2003: 240-247.

[41] Henderson J. and Lemon O. Mixture model POMDPs for efficient handling of uncertainty in dialogue management. //Proceedings of the Annual Meeting of the Association for Computational Linguistics(ACL). Columbus, USA, 2008: 73-76.

[42] Bohus D. and Rudnicky A. A 'k hypotheses + other' belief updating model. //Proceedings of the AAAI Conference on Artificial Intelligence(AAAI). Boston, USA, 2006.

[43] Thomson B., Schatzmann J., and Young S. Bayesian update of dialogue state for robust dialogue systems. //Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP). Las Vegas, USA, 2008: 4937-4940.

[44] Thomson B. and Young S. Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. Computer Speech and Language, 2010, 24(4): 562-588.

[45] Bui T., Poel M., Nijholt A., and Zwiers J. A tractable hybrid DDN-POMDP approach to affective dialogue modeling for probabilistic frame-based dialogue systems. Natural Language Engineering, 2009, (2): 273-307.

[46] Williams J. D. Using particle filters to track dialogue state. //Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding(ASRU). Kyoto, Japan, 2007: 502-507.

[47] Black A. W., Burger S., Conkie A., Hastie H., Keizer S., Lemon O., Merigaud N., Parent G., Schubiner G., Thomson B., Williams J. D., Yu K., Young S., and Eskenazi M. Spoken dialog challenge 2010: Comparison of live and control test results. //Proceedings of Special Interest Group on Discourse and Dialogue(SigDial). Portland, USA, 2011: 2-7.

[48] Williams Jason, Raux Antoine, Ramachandran Deepak, and Black Alan. The dialog state tracking challenge. //Proceedings of Special Interest Group on Discourse and Dialogue(SigDial). Metz, France, 2013.

[49] Jurcicek F., Thomson B., and Young S. Natural actor and belief critic: Reinforcement algorithm for learning parameters of dialogue systems modelled as POMDPs. ACM Transactions on Speech and Language Processing, 2011, 7(3): 6.

[50] Hansen E. Solving POMDPs by searching in policy space. //Proceedings of the Conference on Uncertainty in Artificial Intelligence(UAI). Madison, USA, 1998: 211-219.

[51] Littman M. L., Sutton R. S., and Singh S. Predictive representations of state. //Proceedings of the Annual Conference on Neural Information Processing Systems(NIPS). Vancouver, Canada, 2002, 14: 1555-1561.

[52] Kaelbling L., Littman M., and Cassandra A. Planning and acting in partially observable stochastic domains. Artificial Intelligence, 1998, 101(1): 99-134.

[53] Pineau J., Gordon G., and Thrun S. Point-based value iteration: An anytime algorithm for POMDPs. //Proceedings of the International Joint Conferences on Artificial Intelligence(IJCAI). Acapulco, Mexico, 2003, 3: 1025-1032.

[54] Williams J. and Young S. Scaling pomdps for spoken dialog management. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(7): 2116-2129.

[55] Williams Jason. The best of both worlds: Unifying conventional dialog systems and POMDPs. //Proceedings of the Annual Conference of the International Speech Communication Association(InterSpeech). Brisbane, Australia, 2008: 1173-1176.

[56] Lison P. Towards relational POMDPs for adaptive dialogue management. //Proceedings of the Annual Meeting of the Association for Computational Linguistics(ACL). Uppsala, Sweden, 2010: 7-12.

[57] Kurniawati H., Hsu D., and Lee W. Sarsop: efficient point-based pomdp planning by approximating optimally reachable belief spaces. Proceedings of Robotics: Science and Systems, 2008, 2008: 65-72.

[58] Lefevre F., Gasic M., Jurcicek F., Keizer S., Mairesse F., Thomson B., Yu K., and Young S. k-nearest neighbor monte-carlo control algorithm for pomdp-based dialogue systems. //Proceedings of Special Interest Group on Discourse and Dialogue(SigDial). London, UK, 2009: 272-275.

[59] Schatzmann Jost, Weilhammer Karl, Stuttle Matt, and Young Steve. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. The Knowledge Engineering Review, 2006, 21(2): 97-126.

[60] Getze J., Scheffler Tatjana, Roller Roland, and Reithinger Norbert. User simulation for the evaluation of bus information systems. //Proceedings of the IEEE Spoken Language Technology Workshop(SLT). Berkeley, USA, 2010: 454-459.

[61] Jung Sangkeun, Lee Cheongjae, Kim Kyungduk, Jeong Minwoo, and Lee Gary Geunbae. Data-driven user simulation for automated evaluation of spoken dialog systems. Computer Speech and Language, 2009, 23(4): 479-509.

[62] Schatzmann Jost and Young Steve. The hidden agenda user simulation model. IEEE Transactions on Audio, Speech, and Language Processing, 2009, 17(4): 733-747.

[63] Eckert Wieland, Levin Esther, and Pieraccini Roberto. User modeling for spoken dialogue system evaluation. //Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding. Santa Barbara, USA, 1997: 80-87.

[64] Scheffler Konrad and Young Steve. Probabilistic simulation of human-machine dialogues. //Proceedings of the IEEE International Conference on Acoustics. Speech, and Signal Processing(ICASSP). Istanbul, Turkey, 2000, 2: II1217-II1220.

[65] Scheffler Konrad and Young Steve. Automatic learning of dialogue

strategy using dialogue simulation and reinforcement learning. //Proceedings of the second international conference on Human Language Technology Research. San Diego, USA, 2002: 12-19.

[66] Pietquin Olivier. Framework for Unsupervised Learning of Dialogue Strategies[PhD thesis]. Belgium: Faculte Polytechnique de Mons, 2004.

[67] Chandramohan Senthilkumar, Geist Matthieu, Lefevre Fabrice, Pietquin Olivier, et al. User simulation in dialogue systems using inverse reinforcement learning. //Proceedings of the 12th Annual Conference of the International Speech Communication Association. Florence, Italy, 2011: 1025-1028.

[68] Schatzmann Jost, Georgila Kallirroi, and Young Steve. Quantitative evaluation of user simulation techniques for spoken dialogue systems. //Proceedings of Special Interest Group on Discourse and Dialogue(SigDial). Lisbon, Portugal, 2005: 45-54.

[69] Pietquin Olivier, Hastie Helen, et al. A survey on metrics for the evaluation of user simulations. Knowledge Eng. Review, 2013, 28(1): 59-73.

[70] Zukerman Ingrid and Albrecht David W. Predictive statistical models for user modeling. User Modeling and User-Adapted Interaction, 2001, 11(1-2): 5-18.

[71] Yu Kai. Adaptive training for large vocabulary continuous speech recognition[PhD thesis]. Cambridge, UK: Cambridge University, 2006.

[72] Bellegarda Jerome R. Statistical language model adaptation: review and perspectives. Speech Communication, 2004, 42(1): 93-108.

[73] Yamagishi J. Average-Voice-Based Speech Synthesis[PhD thesis]. Tokyo, Japan: Tokyo Institute of Technology, 2006.

[74] Yu K., Zen H., Mairesse F., and Young S. Context adaptive training with factorized decision trees for hmm-based statistical parametric speech synthesis. Speech Communication, 2011, 53(6): 914-923.

[75] Gasic Milica, Jurcicek Filip, Thomson Blaise, Yu Kai, and Young Steve. On-line policy optimisation of spoken dialogue system via live interaction with human subjects. //Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding(ASRU). Waikoloa, USA, 2011.

[76] Reiter Ehud. Nlg vs. templates. arXiv preprint cmp-lg/9504013, 1995.

[77] Matthiessen Christian MIM. Systemic grammar in computation: the nigel case. //Proceedings of the first conference on European chapter of the Association for Computational Linguistics, Pisa, Italy, 1983: 155-164.

[78] Bateman John A, Maier Elisabeth A, Teich Elke, and Wanner Leo. Towards an architecture for situated text generation. //Proceedings of the International Conference on Current Issues in Computational Linguistics, Penang, Malaysia, 1991: 336-349.

[79] Langkilde Irene and Knight Kevin. Generation that exploits corpus-based statistical knowledge. //Proceedings of the 17th international conference on Computational linguistics Montreal, Canada, 1998, 1: 704-710.

[80] DiMarco Chrysanne. Natural language generation – a survey. 2010. https://cs.uwaterloo.ca/~jchampai/CohenClass.en.pdf

[81] Zhao Yan-yan, Qin Bin and Liu Ting. Text sentiment analysis. Journal of Software, 2010, 21(8): 1834-1848(in Chinese)
(赵妍妍，秦兵，and 刘挺．文本情感分析．软件学报，2010，21(8): 1834-1848)

[82] Wu Chung-Hsien, Hsia Chi-Chun, Lee Chung-Han, and Lin Mai-Chun. Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(6): 1394-1405.

[83] Yu Lin-Li, Cai Zi-Xing and Chen Ming-Yi. Review of sentiment characteristics analysis and recognition research on speech signal. Journal of Circuits and Systems. 2007, 12(4):76-83.(in Chinese)
(余伶俐，蔡自兴，and 陈明义．语音信号的情感特征分析与识别研究综述．电路与系统学报, 2007, 12(4):76-83)

[84] Jurcicek F., Keizer S., Gasic M., Mairesse F., Thomson B., Yu K., and Young S. Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. //Proceedings of the Annual Conference of the International Speech Communication Association(InterSpeech). Florence, Italy, 2011, 11.

[85] Walker M., Litman D., Kamm C., and Abella A. Paradise: A framework for evaluating spoken dialogue agents. //Proceedings of the Annual Meeting of the Association for Computational Linguistics(ACL). Madrid, Spain, 1997: 271-280.

[86] Martin J.C., Abrilian S., Devillers L., Lamolle M., Mancini M., and Pelachaud C. Levels of representation in the annotation of emotion for the specification of expressivity in ECAs. //Proceedings of the International Conference on Intelligent Virtual Agents(IVA). Kos, Greece, 2005: 405-417.

[87] Tato R. S., Kompe R., and Pardo J. M. Emotional space improves emotion recognition. //Proceedings of the International Conference on Spoken Language Processing(ICSLP). 2002.

[88] Cowie R., Cowie E. Douglas, Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., and Taylor J. G. Emotion recognition in human-computer interaction. IEEE Signal Processing magazine, 2001, 18(1): 32-80.

[89] Guven E. Speech emotion recognition using a backyard context. //Proceedings of the Applied Imagery Pattern Recognition Workshop(AIPR). Washington, USA, 2010: 1-5.

[90] Jia Jia, Zhang Shen, Meng Fan-Bo, Wang Yong-Xin, and Cai Lian-Hong. Emotional audio-visual speech synthesis based on pad. IEEE Transactions on Audio, Speech and Language Processing, 2011, 19(3): 570-582.

[91] Zhang Shen, Xu Yingjin, Jia Jia, and Cai Lian-Hong. Analysis and

modeling of affective audio visual speech based on PAD emotion space. //Proceedings of the 6th International Symposium on Chinese Spoken Language Processing(ISCSLP). Kunming, China, 2008: 1-4.

[92] Tanenhaus Michael K and Brown-Schmidt Sarah. Language processing in the natural world. The Royal Society. 2008, 363(1493): 1105-1122.

**YU Kai**, born in 1976, P.h.D., research professor. His research interest includes cognitive spoken dialogue system, speech synthesis, recognition, understanding and machine learning.

**CHEN Lu**, born in 1990, Ph.D. Candidate. His research interest includes statistical spoken dialogue systems and machine learning.

**CHEN Bo**, born in 1990, Postgraduate . His research interest includes affective interaction and machine learning.

**SUN Kai**, born in 1992, Undergraduate. His research interest includes human-machine interaction, machine learning and human-machine chess.

**ZHU Su**, born in 1990, Postgraduate. His research interest includes speech understanding, natural language processing and machine learning.



**Background**

This paper reviews the cognitive technology in human machine dialogue systems, in particular, spoken dialogue systems. It is a disciplinary topic covering human-machine interaction, cognitive engineering and intelligent speech technology. The key issue to address is how to build a cognitive agent for task-oriented conversational human-machine interaction. Although general spoken dialogue systems and speech recognition have been studied for years, cognitive technology just emerged as a new attractive research direction due to the advances of computing hardware and software as well as the booming of mobile internet. It enables the machine to be capable of understanding, learning, guiding and adapting. Deep and robust understanding, inference based on uncertain information, policy optimization, adaptation and influential information generation are the main research areas. This paper is a position paper of cognitive technology in spoken dialogue systems. We argue that task-oriented dialogue system consists of three layers: physical layer, control layer and application layer. IO technology, cognitive technology and knowledge management are corresponding techniques. As far as we know, this is the first work to introduce and define a independent "cognitive control layer" for human machine interaction. The scope and content is introduced in detail. Relevant emerging techniques are reviewed and future research direction are also discussed. International researchers have performed small scale research on the above research topics individually. However, systematic research and research on large-scale tasks or big data are still open problems.

The SpeechLab at SJTU focus on the research of large-scale real world human machine spoken interactive systems. A number of interesting research results have been achieved, including fast and parallel training of large scale speech recognition, statistical speech understanding, robust tracking of dialogue states and the building of real world speech recognition and spoken dialogue systems. This paper was supported by the NSFC excellent young researcher project "human-machine spoken dialogue systems", which aims at advancing the research of cognitive spoken dialogue systems.