自然语言对话关键技术及系统

尚利峰 蒋 欣 陈 晓
华为诺亚方舟实验室

关键词：自然语言对话系统 深度学习

自然语言对话是网络大数据语义理解的主要挑战之一，被誉为人工智能皇冠上的宝石。近年来随着深度学习的发展，无论是以情感交流为目的的闲聊式对话机器人，还是以帮助人们完成各种不同任务的任务型对话机器人，都已经开始走进并逐步融入到我们的生活中。自然语言对话系统研究的目的是希望机器人能够理解人类的自然语言，同时进行个性化的情感表达、知识推理、信息汇总等任务。本文回顾了自然对话系统近二三十年来的研究工作，从传统的基于概率决策过程的多轮对话系统，到近年来提出的基于深度学习的生成式对话系统，再到将深度学习和符号处理相融合的神经符号对话系统。最后我们指出，更好地进行自然语言理解、知识表示和推理是整个对话系统发展的核心推动力。

自然语言对话的发展

网络大数据中的语义理解，是让机器理解人类的自然语言，有两种不同的定义[19]。一种是基于表示的 (representation-based)，系统能够根据输入的自然语言产生合理的内部表示。例如，输入“姚明的身高”，系统能够联系到篮球运动员“姚明”以及他的属性“身高”，就意味着机器理解了语言。另一种是基于行为的 (behavior-based)，系统根据输入的语言采取正确的动作。例如，有人说“打开电视并收看央视四套”，机器正确执行了这个命令，就认为它理解了该语言。

对话是指两人或者多人之间进行的交谈，在日常工作和学习中人们通过对话获取信息、交流情感、表达观点或者寻求帮助。研究自然语言对话的目的是希望构建一个对话机器人，它可以像人一样进行多轮对话，帮助我们完成不同的任务。对话系统（或对话机器人）也是一种新的人机交互方式，被认为是智能设备最重要的交互方式之一。典型的对话系统如苹果 Siri、微软小冰、亚马逊 Alexa 等，已经成为人们日常生活的重要助手。商用的对话系统目前还是由多个垂直领域机器人构成的综合体，每个领域的机器人只涵盖特定的应用场景，比如特定任务（如订机票）、闲聊、百科问答等。要构建一个具备物理和社会常识、能够理解自然语言、会自由表达情感的通用对话机器人依然充满很多技术挑战。

自然语言对话研究的挑战主要来自于对话的三个基本特性：(1) 多样性：参与对话的个体有独特的理解、思考和表达方式；(2) 连贯性：多轮对话的内容前后是有关联和逻辑的；(3) 知识性：参与对话的个体有常识和特定专业知识，通过对话传递知识或情感。

表达的多样性为自然语言的表示和理解带来了很大的挑战。举例来说，想通过手机语音助手连接 Wi-Fi 网络，最直接的表达方式是“打开 Wi-Fi”，而有些人则会根据自身的生活经验说“我想用家里的网络”。这两种表达方式完全不同，如何让机器人知道它们表达了同样的意图？对话系统有一个自然语言理解模块，可将这些不同的表达方式转换成同一个语义表示。转换过程通常需要依赖背景知识

或者常识（比如：家里有宽带上网）。如何对知识和常识进行表示和推理是对话系统研究的另外一个核心问题。目前大部分的对话机器人只能进行单轮对话（比如控制机器人调节音箱的音量）。为了构建多轮对话系统，需要专门的模块来管理历史对话状态，并且每轮对话的解析都需要考虑历史对话状态，因此对话的多轮属性为自然语言理解带来了更大的挑战。在理解自然语言的基础上，对话机器人通过执行特定的操作或者查询知识库帮助用户完成任务或回答问题。

自然语言理解、知识表示和推理的研究推动了整个对话系统的发展。第一代对话系统可以追溯到1980年代末发展起来的基于专家制定的语法规则和本体设计的规则系统。第二代系统是在1990年代发展起来的数据驱动的统计对话系统，同时在这一时期强化学习方法也被引入到对话管理当中。近年发展起来的基于深度学习的端到端对话系统可以认为是第三代技术。不同的对话技术并不是割裂互斥的，第三代技术并不是对前面两代技术的完全取代。在商业系统中可能需要将各代技术相结合，实现其优势互补。例如，如何将深度学习和基于符号处理的第一代技术结合起来是当前业界非常关注的一个研究课题。

对话系统组成模块和基本原理

自然语言对话系统主要由语音识别、自然语言理解、对话管理、自然语言生成和语音合成五个基本模块构成（如图1所示）。

语音识别是将语音信号转换成文本信息。目前，在远场（说话人距离麦克风较远）和有噪音的情况下，语音识别正确率仍然不高，这也为后面的自然语言理解带来了一些挑战。为此，需要针对场景进行语音识别的定制优化，或者开发相应的纠错模块。

自然语言理解模块将文本信息转换成结构化的语义表示。自然语言的表达非常多样，同一个意图可以有很多不同的表达方式，比如“怎么截取手机屏幕？”和“如何像电脑一样获取手机桌面？”都表达了截屏的诉求，但是字面上差别很大。“打电话”这样简单的一个意图就至少有上百种不同的表达方式。除了输入本身，自然语言的理解还需要依赖上下文、领域知识和常识。自然语言的表示和理解是发展通用对话机器人的核心问题，而当前并不存在一个“魔法盒子”可以实现通用的自然语言理解。

对话管理由两个子模块构成：对话状态更新和行为选择。对话状态更新基于历史对话状态和当前用户输入。行为选择则是基于当前的对话状态决定机器人的下一步行为。在行为选择的过程中除了需要考虑业务逻辑的限制，通常还需要基于外部知识库的查询结果。比如，用户说“帮我订一张明天下午从北京到上海的机票”，对话管理模块需要根据机票数据库的查询结果来决定是订购机票还是告知用户无票。对话状态更新和行为选择之间可能是交替进行的。对话管理本质上是一个决策过程，可以通过人工编写规则来实现，也可以通过数据驱动的统计方法来实现。

自然语言生成是将机器人的行为转换成自然语言文本，常见的方式有基于规则、模板填充和深度学习的方法。语音合成则是将自然语言文本转化为语音输出。

需要指出的是，并不是所有的对话系统都严格按照图1所示的流程进行搭建，特别是随着深度学习的发展，有人提出来可以将语音识别和自然语言理解合并成一个“端到端”的模型，将语音输入直接转换成语义输出。

统计对话系统

基于部分可观测马尔科夫决策过程(Partially Observable Markov Decision Processes, POMDP)的统计对话系统[1]（如图2所示）将用户看作是机器人的外部环境，自然语言理解模块是机器人的感知系统，o_t表示对环境的（有噪声的）感知，机器人根据置信状态b_t选择行为a_t并执行，最终用户对

机器人的行为进行评价得到 r_t。其中置信状态 b_t 表示所有可能内部状态 s_t 的分布。对话管理是通过马尔科夫决策过程来实现的，由对话模型和策略模型构成。对话模型对应图 1 中的状态更新模块，策略模型则与行为选择相对应。对话模型主要涉及到两个概率分布：观测概率 $P(o_t|s_t, a_{t-1})$ 和转移概率 $P(s_t|s_{t-1}, a_{t-1})$。通过这两个概率，基于当前的用户输入 o_t 和历史对话状态 s_{t-1}，共同约束限定了 s_t 的分布。策略模型则是关于置信状态 b_t 和对话行为 a_t 的函数。整个对话系统的参数可以通过监督学习和强化学习来估计。

在基于 POMDP 的统计对话系统中，对话的内部状态不再是某一个确定的状态 s_t，而是考虑了所有可能对话状态分布的置信状态 b_t。用户的输入也不是某一个确定的语义解析，而是关于用户行为的分布。因此该方法考虑了所有可能的内部对话状态的转移路径，从而对输入噪声有更好的鲁棒性。

深度对话系统

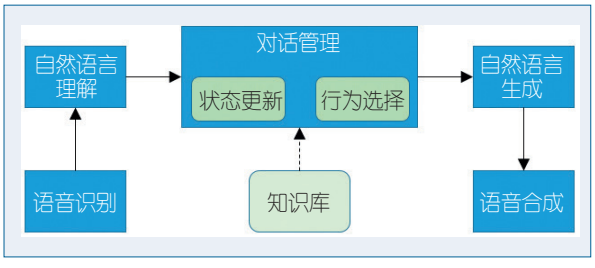

图1　语音对话系统流程图

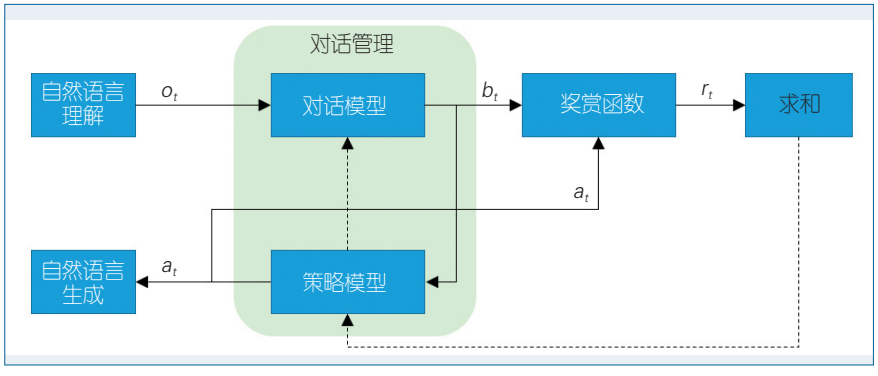

图2　基于POMDP的统计对话系统[1]

深度学习为自然语言处理带来的关键性突破是语句的语义表示学习[2-4]。当自然语言的语义被表示为高维向量，整个对话系统也可以基于这些语义向量进行构建，然后通过反向传播算法 (backpropagation) 对模型的参数进行学习。深度学习最开始被用于闲聊式对话中，这种类型的对话不涉及显式的知识和业务逻辑约束，比较容易通过端到端的深度模型建模，后来则逐步发展到包含知识的问答和任务驱动的多轮对话系统中。

检索式对话系统

传统的闲聊式对话是基于规则和检索来实现的[5]。基于检索的对话系统没有独立的对话管理和自然语言生成模块，主要是通过自然语言理解模块将输入文本表示为语义向量，然后通过查询对话知识库找到最优的回复，所以其核心的研究问题是如何表示文本和度量文本之间的语义相关性。相关性计算方法除了经典的基于余弦相似度、话题、翻译等模型外[6]，深度匹配模型也被应用到对话系统中。

卷积神经网络 (Convolutional Neural Networks, CNNs) 是搭建计算机视觉的核心组件。Hu 等人首先把卷积神经网络用在深度文本匹配中，并提出了两个匹配框架以度量文本在不同尺度上的匹配[7,8]。框架一的神经网络模型如图 3 所示。对于输入的两个文本，首先通过使用卷积神经网络分别得到两个文本的语义向量表示，然后将这两个语义向量连接在一起作为后续多层感知机 (Multi-Layer Perceptron, MLP) 的输入，最终输出两个输入文本的语义相关度。该框架的不足之处是匹配发生在两个句子的全局语义表示上，其中局部的匹配特征可能被忽略。

为了克服框架一的不足，框架二（如图 4 所示）在第一层考虑了输入文本之

间所有可能的局部语义匹配表示的组合，基于这些局部匹配组合，二维卷积和池化层抽取出所有可能的局部匹配特征，最后通过多层感知机得到文本的语义相关性。实验表明框架二比框架一有更好的效果，但是模型的复杂度和计算量也相对更大。

基于卷积神经网络的深度匹配模型并没有考虑文本的句法结构信息，Wang 等人提出基于句法结构的深度匹配模型，对输入文本进行依存句法分析，并从海量的句法树对中挖掘匹配模式，将匹配模式输入深度神经网络模型，学习短文本的匹配关系，取得了较好的效果[9]。

生成式对话系统

在现实世界中，由于语言的复杂性和多样性，很难通过基于规则的方法来构建一个开放领域的对话系统，因为规则冲突和爆炸的难题不可避免。另一方面，相比于人类所有可能的对话，我们通常只能获取到一个小而稀疏的对话数据集合。因此在基于检索的对话系统中，无论我们定义多么智能的匹配模型，都无法应对所有可能的对话情形。因此我们需要一种新的架构——生成式对话模型，该模型可以在理解人类对话的基础上，自适应地根据对话内容生成新的对话[10~12]。

以文献 [10] 中提出的神经响应机 (neural responding machine) 为例。神经响应机主要由编码器和解码器两个部分组成（如图 5 所示）。编码器实现“语义理解”的功能，解码器则实现“语言生成”的功能。在图 5 的例子中，输入文本“华为手机怎么样”，可以认为编码器是对输入文本信息在不同的尺度上进行信息的表示和记忆。具体来讲，该模型采用了两种不同的编码器：局部编码器和全局编码器。全局编码器实现对句子整体语义的理解，局

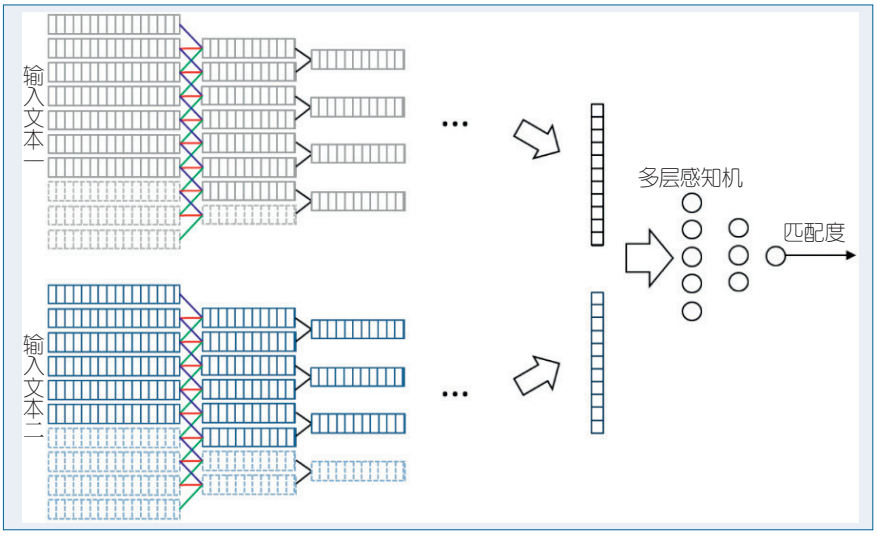

图3 基于卷积神经网络的深度匹配模型-框架一[7]

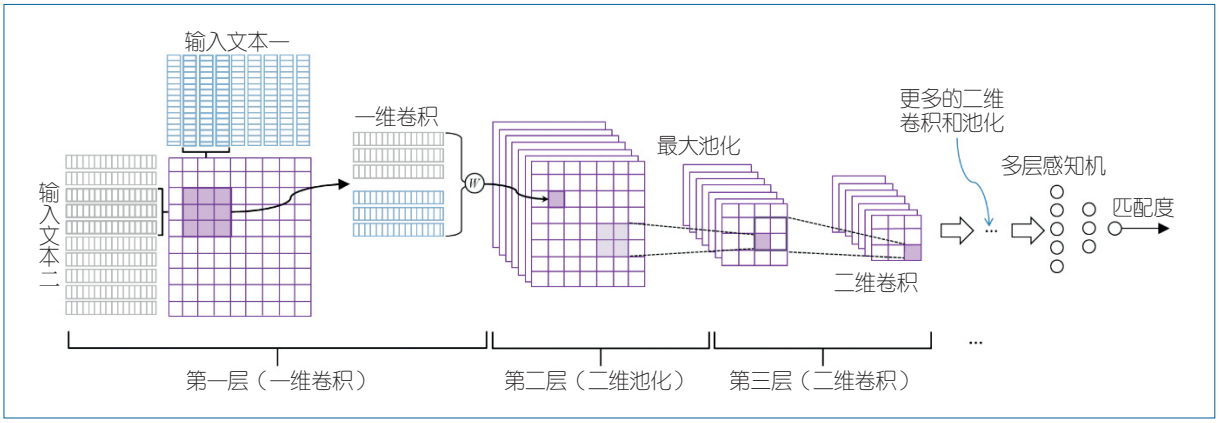

图4 基于卷积神经网络的深度匹配模型-框架二[7]

部编码器对句子的各个局部语义进行表示。就像人类在思考如何进行对话时，除了需要把握整句话的语义，还需要对某些局部语义进行特别的思考和关注，以便进行针对性的回复。基于对输入文本的编码，解码器部分就可以自适应地根据对输入文本的局部和全局的理解，逐词地生成回复“用了都说好”。神经响应机主要用来进行单轮对话，至于多轮对话则需要将全局编码器替换为一个分层神经网络以对历史对话进行建模。

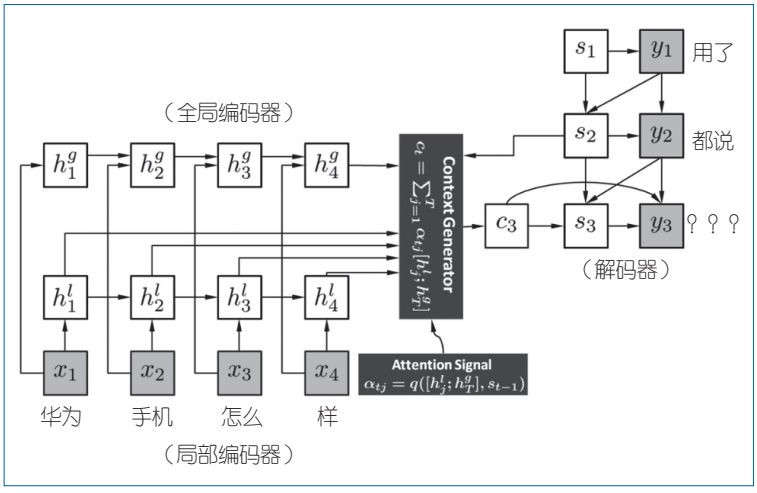

图5　神经响应机

对话和翻译都可以通过类似的序列到序列(sequence-to-sequence)模型来建模。不同之处在于，同一句话的不同翻译基本保持同样的语义，字面上也不会有太大的差别。而在对话中，对于同一个用户输入可以有很多语义不一样的回复。因此，对话模型需要对各种可能的回复进行表达，注意力机制(attention mechanism)的引入有效提升了对话系统的表达能力[13]。如果将神经响应机放到图1所示的框架下来考虑的话，注意力机制则实现对话管理模块的部分功能，该机制动态地对用户的输入给出不同的理解，由此生成不同的回复。

有“复制”机制的生成式对话系统

在实际对话中，我们经常需要在回复中“复制”输入的部分文本。比如用户说“青岛四季宜人”，神经响应机可能给出回复“壮哉我大深圳”，这句话虽然语句通顺，但是“答非所问”，不是一个好的回复。相关的回复中应该提到“青岛”，例如“壮哉我大青岛”。

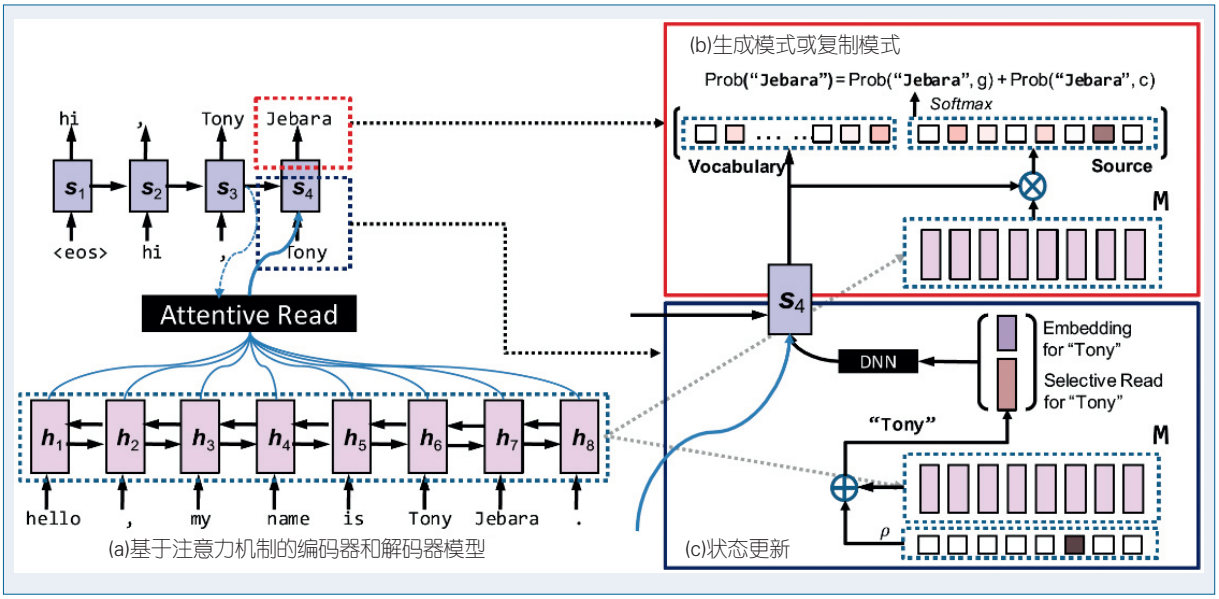

图6　有“复制”机制的对话系统[14]

Gu 等人提出了有“复制”机制的对话系统模型[14]（如图 6 所示）。该模型的解码器部分有“生成”和“复制”两种模式，在生成每一个词时，通过一个混合概率模型来决定选择哪种模式。除了对话，带有复制机制的模型也广泛用在文本摘要任务中，并且已经成为该任务的基准模型。

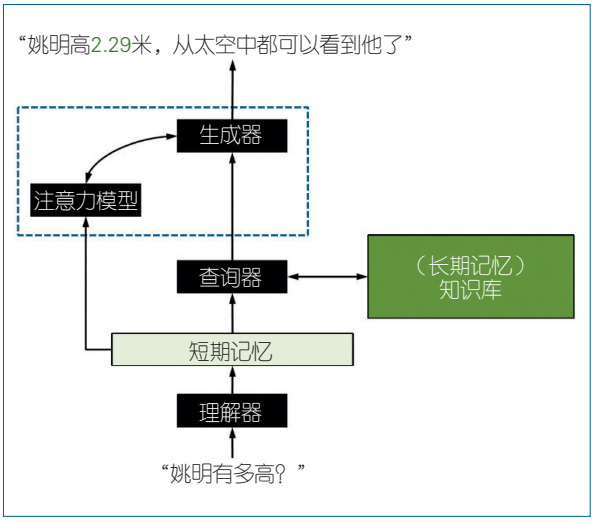

图7 神经生成式问答系统[15]

连接知识库的生成式对话系统

传递知识是对话中非常重要的一个功能。例如，用户问“姚明有多高呢？”，回复可能是“他很高”，但这样的回复没有包含具体的知识。如果回复是“2.29米”，虽然事实正确，但缺少对话中应有的流畅的衔接。为了生成既包含正确的知识又流畅的回复，Yin 等人提出了一种可以查询外部知识库的端到端深度对话系统：神经生成式问答系统[15]（如图 7 所示）。该系统包含了一个新的模块“神经查询器”，负责知识库的查询。查询的结果是知识库中和用户问题最相关的知识点。其中，问题和知识点的相关性是通过匹配模型来实现的，基于卷积神经网络的深度匹配表现最优。另外一个改变是在“生成器”部分，通过一个混合概率模型，判断当前生成的词是否来自于查询到的相关知识点。实验表明该对话系统在回复提问的效果上超过了传统的基于检索的问答系统及神经响应机。

该模型中的知识点是以三元组的形式存在的，比如：（姚明，身高，2.29 米），三元组的内容可以通过语义向量来表示，因此整个模型都是基于向量表示来构建的，可以通过反向传播算法进行端到端的学习。

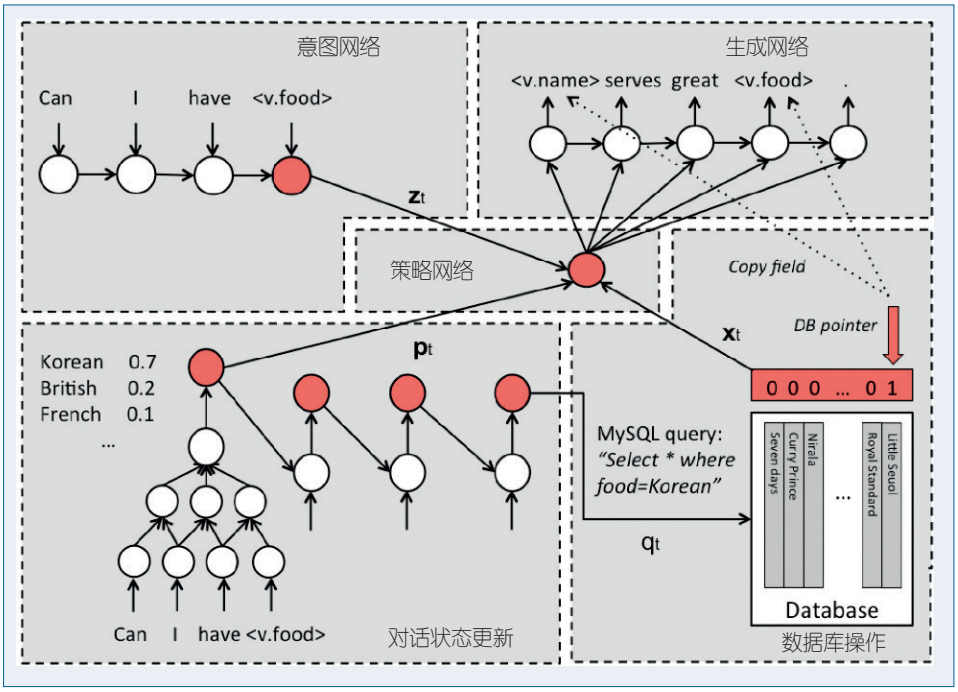

图8 任务驱动的多轮深度对话系统[16]

任务驱动的多轮深度对话系统

上述深度对话模型没有显式的对话管理模块，不太适合作为以任务完成为目的的多轮对话系统。Wen 等人提出了一种以任务完成为驱动的多轮深度对话系统（如图 8 所

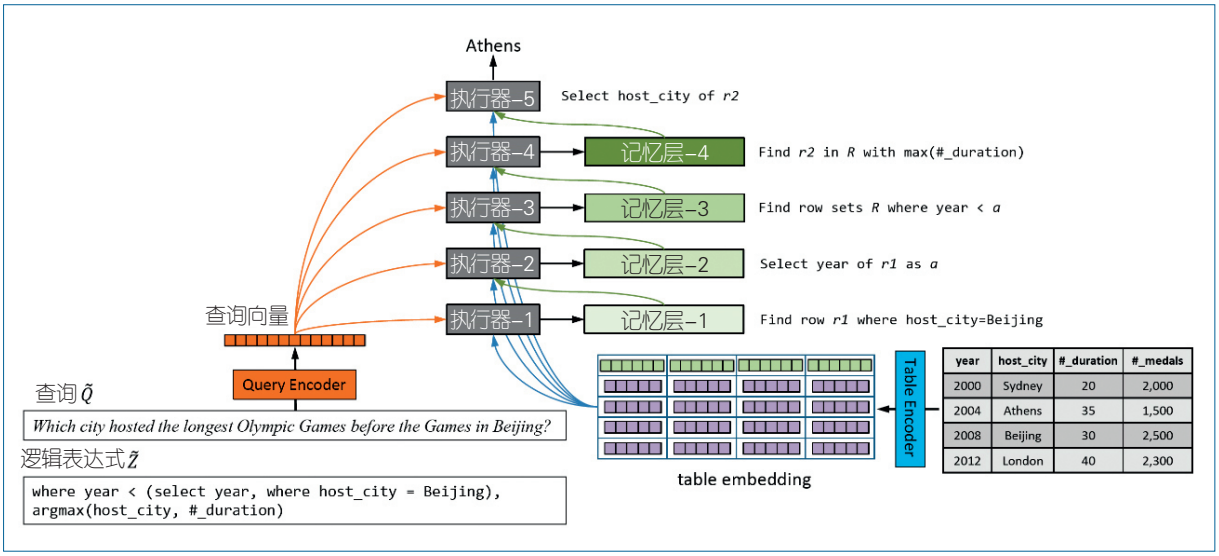

图9 神经查询器[17]

示），该系统包含了标准对话系统的所有模块，每个模块都是通过神经网络进行建模[16]。其中，自然语言理解部分通过循环神经网络（或者卷积神经网络）进行用户意图识别，对话管理中的状态更新模块是通过乔丹 (Jordan) 循环神经网络来实现的，策略选择部分的输入来自于更新后的对话状态、数据库的查询结果和用户当前输入的意图向量，这些输入通过一个深度网络计算得到机器的行为。自然语言生成部分则是另外一个循环神经网络，基于机器行为得到最终的自然语言输出。

该模型的训练分成两个阶段。首先训练对话状态更新模块，保证每次都可以正确地解析业务相关的槽位 (slot) 信息。然后，固定对话状态更新模块的参数，训练其他所有模块的参数。事实上，该模型并不是一个严格意义上可以进行端到端训练的模型，主要是因为数据库查询是通过符号化的 SQL 语句来实现的，无法直接进行梯度计算。

为了解决数据库访问无法进行梯度计算的问题，Yin 等人提出了神经查询器（如图 9 所示），该模型首先提出将整个数据库用向量来表示，对数据库表里面的每个元素进行表示学习，通过这种方式，整个模型就可以进行端到端的训练[17]。

虽然深度学习可以很好地应对语言多样性，但是符号化的表示和处理在业务逻辑、知识查询和推理等方面更加高效，有更好的可解释性。因此，如何对深度学习和符号处理进行融合是一个很重要的研究课题。Mou 等人提出的连接查询器 (coupled enquirer) 是在这个方向上的一个重要探索[18]。该工作将神经网络操作和离散的符号操作相结合，通过逐层的神经网络执行器 (neural executor) 的中间结果，对符号执行器 (symbolic executor) 进行强化学习。同时，一个经过更好训练的符号执行器，能够进一步提升神经网络执行器的效果。对于深度学习技术和符号处理的结合，未来还有很大的研究空间。

结束语

从自然语言对话系统的发展历程中可以看到，自然语言理解，特别是结合上下文对话历史的自然语言理解，依然是对话研究的关键难题。大数据与深度学习技术的进步为自然语言对话研究提供了许多新的工具和方法，大量新的模型如雨后春笋般被研究出来，基于深度学习的闲聊式对话已经在实际产品中应用。然而，任务型对话系统的可扩展性依然不够理想，在一个垂直领域开发的自然语言理解模块很难快捷地复用到其他领域，可迁移的语义表

示和理解方法仍然是我们亟需攻克的难题。 ■

尚利峰
华为诺亚方舟实验室研究员。主要从事对话系统方面的研究和开发。
Shang.Lifeng@huawei.com

蒋　欣
华为诺亚方舟实验室主任研究员。主要研究方向为自然语言处理、信息检索和机器学习。
Jiang.Xin@huawei.com

陈　晓
华为诺亚方舟实验室研究员。主要研究方向为自然语言基本问题（分词、词性标注、句法分析、语义分析）、机器翻译、语音识别、对话系统等。
chen.xiao2@huawei.com

参考文献

[1] Young S, Gašić M, Thomson B, et al. POMDP-Based Statistical Spoken Dialog Systems: A Review[J]. *Proceedings of the IEEE*, 2013, 101(5):1160-1179.

[2] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]// *INTERSPEECH 2010, Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, September. DBLP, 2010:1045-1048.

[3] Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings[J]. *TCAL*, 2015, 3:211-225.

[4] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. *Advances in Neural Information Processing Systems*, 2013, 26:3111-3119.

[5] Ji Z, Lu Z, Li H. An Information Retrieval Approach to Short Text Conversation[J]. *Computer Science*, 2014.

[6] Li H, Xu J. Semantic Matching in Search[J]. *Foundations and Trends in Information Retrieval*, 2014, 7(5):343-469.

[7] Hu B, Lu Z, Li H, et al. Convolutional Neural Network Architectures for Matching Natural Language Sentences[J]. *Advances in Neural Information Processing Systems*, 2015, 3:2042-2050.

[8] Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences[J]. Eprint Arxiv, 2014, 1.

[9] Mingxuan Wang, Zhengdong Lu, Hang Li, Qun Liu. Syntax-based Deep Matching of Short Texts. IJCAI'15, 2015. Wang M, Lu Z, Li H, et al. Syntax-Based Deep Matching of Short Texts.[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2015:1354-1361.

[10]Shang L, Lu Z, Li H. Neural Responding Machine for Short-Text Conversation[J]. 2015:52-58.

[11]Vinyals O, Le Q. A Neural Conversational Model[J]. Computer Science, 2015.

[12]Sordoni A, Galley M, Auli M,et al. A neural network approach to context-sensitive generation of conversational responses[J]. NAACL-HLT 2015.

[13]Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.

[14]Gu J, Lu Z, Li H, et al. Incorporating Copying Mechanism in Sequence-to-Sequence Learning[J]. 2016:1631-1640.

[15]Jun Y, Jiang X, Lu Z, et al. Neural generative question answering,2015.

[16]Wen T H, Vandyke D, Mrksic N, et al. A Network-based End-to-End Trainable Task-oriented Dialogue System[J]. 2016.

[17]Yin P, Lu Z, Li H, et al. Neural Enquirer: Learning to Query Tables in Natural Language[C]// The Workshop on Human-Computer Question Answering. 2016:29-35.

[18]Mou L, Lu Z, Li H, et al. Coupling Distributed and Symbolic Execution for Natural Language Queries[J]// Proceedings of the 34th International Conference on Machine Learning .2017.

[19]李航 . 自然语言对话：现状和未来 .2018。