如何成为一名对话系统工程师

`NLP`

目录

对话系统（对话机器人）本质上是通过机器学习和人工智能等技术让机器理解人的语言。它包含了诸多学科方法的融合使用，是人工智能领域的一个技术集中演练营。图1给出了对话系统开发中涉及到的主要技术。

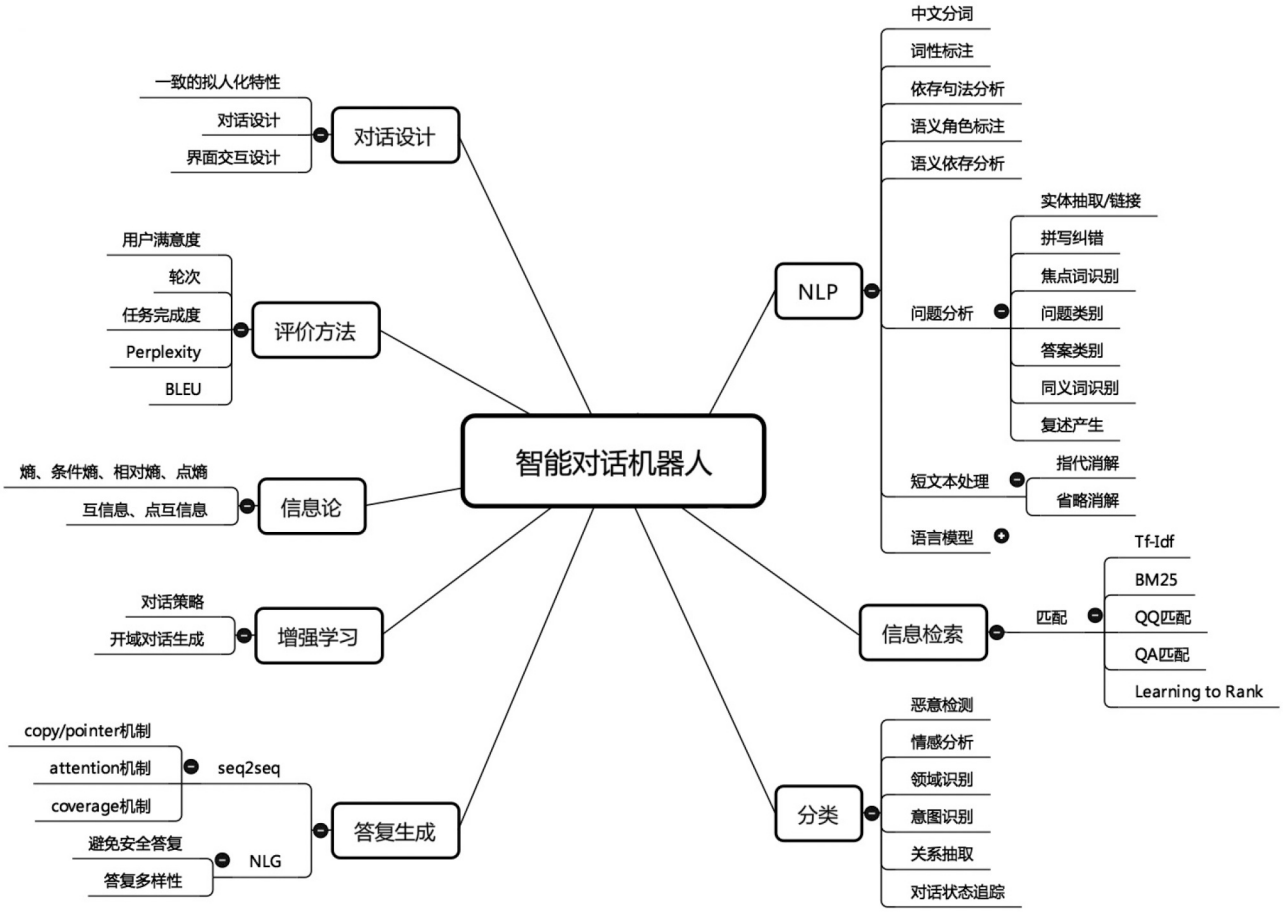

图1　对话系统技能树

对话系统技能进阶之路

图1给出的诸多对话系统相关技术，从哪些渠道可以了解到呢？下面逐步给出说明。

一、对话系统技能进阶之路

1.1 数学

矩阵计算主要研究单个矩阵或多个矩阵相互作用时的一些性质。机器学习的各种模型都大量涉及矩阵相关性质，比如 PCA 其实是在计算特征向量，MF 其实是在模拟 SVD 计算奇异值向量。人工智能领域的很多工具都是以矩阵语言来编程的，比如主流的深度学习框架，如 Tensorflow、PyTorch 等无一例外。矩阵计算有很多教科书，找本难度适合自己的看看即可。如果想较深入理解，强烈推荐《Linear Algebra Done Right》这本书。

概率统计是机器学习的基础。常用的几个概率统计概念：随机变量、离散随机变量、连续随机

变量、概率密度/分布（二项式分布、多项式分布、高斯分布、指指数族分布）、条件概率密度/分布、先验密度/分布、后验密度/分布、最大似然估计、最大后验估计。简单了解的话可以去翻翻经典的机器学习教材，比如《Pattern Recognition and Machine Learning》的前两章，《Machine Learning：A Probabilistic Perspective》的前两章。系统学习的话可以找本大学里概率统计里的教材。

最优化方法被广泛用于机器学习模型的训练。机器学习中常见的几个最优化概念：凸/非凸函数、梯度下降、随机梯度下降、原始对偶问题。一般机器学习教材或者课程都会讲一点最优化的知识，比如 Andrew Ng 机器学习课程中 Zico Kolter 讲的《Convex Optimization Overview》。当然要想系统了解，最好的方法就是看 Boyd 的《Convex Optimization》书，以及对应的 PPT和课程。喜欢看代码的同学也可以看看开源机器学习项目中涉及到的优化方法，例如 Liblinear、LibSVM、Tensorflow就是不错的选择。

常用的一些数学计算 Python 包：

- NumPy：用于张量计算的科学计算包
- SciPy：专为科学和工程设计的数学计算工具包
- Matplotlib：画图、可视化包

1.2 机器学习和深度学习

Andrew Ng 的“Machine Learning”课程依旧是机器学习领域的入门神器。不要小瞧所谓的入门，真把这里面的知识理解透，完全可以去应聘算法工程师职位了。推荐几本公认的好教材：Hastie 等人的《The Elements of Statistical Learning》，Bishop的《Pattern Recognition and Machine Learning》，Murphy 的《MachineLearning：A Probabilistic Perspective》，以及周志华的西瓜书《机器学习》。深度学习资料推荐 Yoshua Bengio 等人的《Deep Learning》，以及 Tensorflow 的官方教程。

常用的一些工具：

- cikit-learn：包含各种机器学习模型的 Python 包
- Liblinear：包含线性模型的多种高效训练方法
- LibSVM：包含各种 SVM 的多种高效训练方法
- Tensorflow：Google 的深度学习框架
- PyTorch：Facebook 的深度学习框架

- Keras：高层的深度学习使用框架
- Caffe：老牌深度学习框架

1.3 自然语言处理

很多大学都有 NLP 相关的研究团队，比如斯坦福 NLP 组，以及国内的哈工大 SCIR 实验室等。这些团队的动态值得关注。

NLP 相关的资料网上随处可见，课程推荐斯坦福的“CS224n：Natural Language Processing with Deep Learning”，书推荐Manning的《Foundations of Statistical Natural Language Processing》（中文版叫《统计自然语言处理基础》）。

信息检索方面，推荐 Manning 的经典书《Introduction to Information Retrieval》（王斌老师翻译的中文版《信息检索导论》），以及斯坦福课程“CS 276：Information Retrieval and Web Search”。

常用的一些工具：

- jieba：中文分词和词性标注 Python 包
- CoreNLP：斯坦福的 NLP 工具（ Java ）
- NLTK：自然语言工具包
- TextGrocery：高效的短文本分类工具（注：只适用于 Python 2）
- LTP：哈工大的中文自然语言处理工
- Gensim：文本分析工具，包含了多种主题模型
- Word2vec：高效的词表示学习工具
- GloVe：斯坦福的词表示学习工具
- Fasttext：高效的词表示学习和句子分类库
- FuzzyWuzzy：计算文本之间相似度的工具
- CRF++：轻量级条件随机场库（C++）
- Elasticsearch：开源搜索引擎

二、对话机器人

对话系统针对用户不同类型的问题，在技术上会使用不同的框架。下面介绍几种不同类型的对

话机器人。

对话机器人创建平台
如果你只是想把一个功能较简单的对话机器人（ Bot ）应用于自己的产品，Bot 创建平台是最好的选择。Bot 创建平台帮助没有人工智能技术积累的用户和企业快速创建对话机器人，国外比较典型的 Bot 创建平台有 Facebook 的 Wit.ai 和 Google 的 Dialogflow（前身为 Api.ai），国内也有不少创业团队在做这方面的事，比如一个 AI、知麻、如意等。

2.1 检索型单轮对话机器人

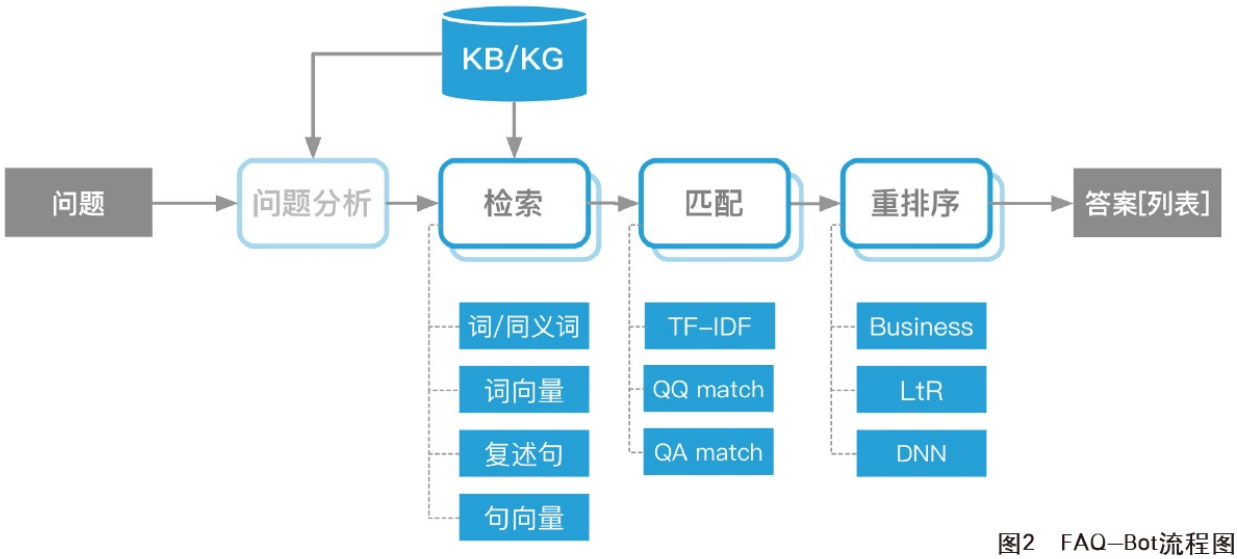

图2　FAQ-Bot流程图

检索型单轮机器人（ FQA-Bot ）涉及到的技术和信息检索类似，流程图2所示。

因为 query 和候选答案包含的词都很少，所以会利用同义词和复述等技术对 query 和候选答案进行扩展和改写。词表示工具 Word2vec、GloVe、Fasttext 等可以获得每个词的向量表示，然后使用这些词向量计算每对词之间的相似性，获得同义词候选集。当然同义词也可以通过已经存在的结构化知识源如 WordNet、HowNet 等获得。复述可以使用一些半监督方法如 DIRT 在单语语料上进行构建，也可以使用双语语料进行构建。PPDB 网站包含了很多从双语语料构建出来的复述数据集。

2.2 知识图谱型机器人

知识图谱型机器人（ KG-Bot，也称为问答系统），利用知识图谱进行推理并回答一些事实型

问题。

知识图谱通常把知识表示成三元组——（主语、关系、宾语），其中关系表示主语和宾语之间存在的某种关系。

构建通用的知识图谱非常困难，不建议从0开始构建。我们可以直接使用一些公开的通用知识图谱，如 YAGO、DBpedia、CN-DBpedia、Freebase 等。特定领域知识图谱的构建可参考“知识图谱技术原介绍”，“最全知识图谱综述#1：概念以及构建技术”等文章。知识图谱可以使用图数据库存储，如 Neo4j、OrientDB 等。当然如果数据量小的话 MySQL、SQLite 也是不错的选择。

为了把用户 query 映射到知识图谱的三元组上，通常会使用到实体链接（把 query 中的实体对应到知识图谱中的实体）、关系抽取（识别 query 中包含的关系）和知识推理（ query 可能包含多个而不是单个关系，对应知识图谱中的一条路径，推理就是找出这条路径）等技术。

2.3 任务型多轮对话机器人

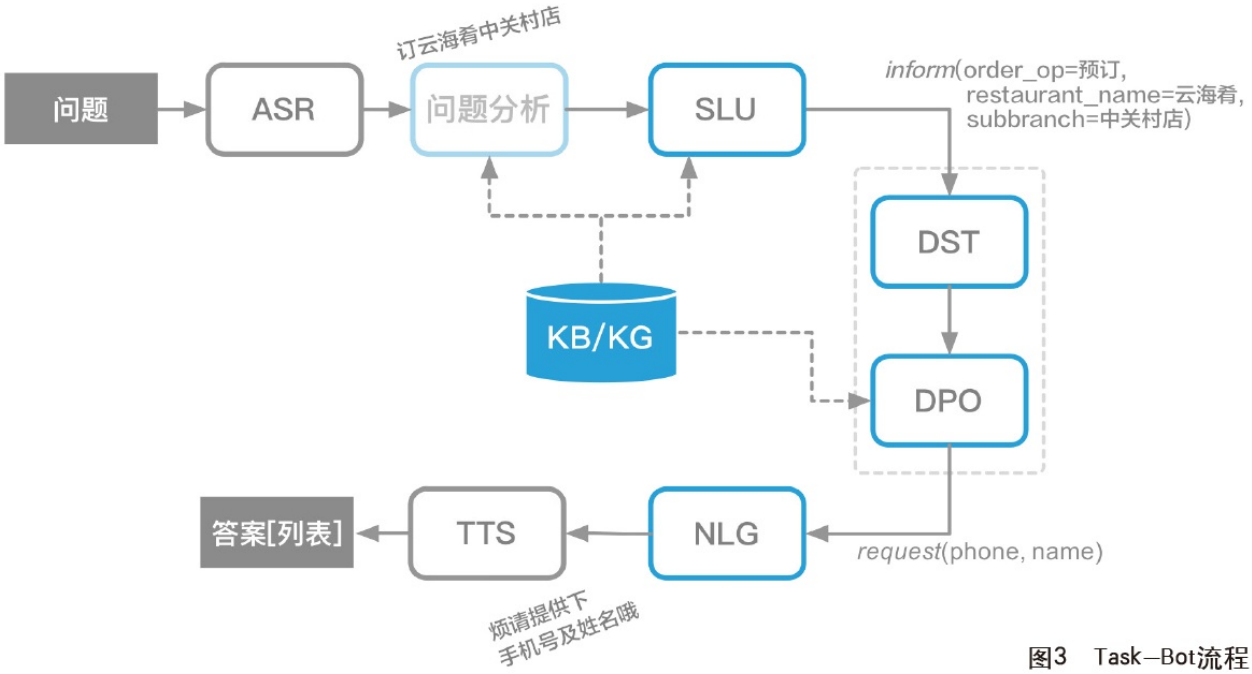

图3　Task—Bot流程图

任务型多轮机器人（ Task-Bot ）通过多次与用户对话交互来辅助用户完成某项明确具体的任务，流程图见图3。

除了与语音交互的 ASR 和 TTS 部分，它包含以下几个流程：

- 语言理解（ SLU ）：把用户输入的自然语言转变为结构化信息—— act-slot-value 三元组。例如餐厅订座应用中用户说“订云海肴中关村店”，我们通过 NLU 把它转化为结构化信息：“inform（ order_op=预订，restaurant_name=云海肴，subbranch=中关村店）”，其中的“inform”是动作名称，而括号中的是识别出的槽位及其取值。

NLU 可以使用语义解析或语义标注的方式获得，也可以把它分解为多个分类任务来解决，典型代表是 Semantic Tuple Classifier（ STC ）模型。

- 对话管理（ DM ）：综合用户当前 query 和历史对话中已获得的信息后，给出机器答复的结构化表示。对话管理包含两个模块：对话状态追踪（ DST ）和策略优化（ DPO ）。

DST 维护对话状态，它依据最新的系统和用户行为，把旧对话状态更新为新对话状态。其中对话状态应该包含持续对话所需要的各种信息。

DPO 根据 DST 维护的对话状态，确定当前状态下机器人应如何进行答复，也即采取何种策略答复是最优的。这是典型的增强学习问题，所以可以使用 DQN 等深度增强学习模型进行建模。系统动作和槽位较少时也可以把此问题视为分类问题。

- 自然语言产生（ NLG ）：把 DM 输出的结构化对话策略还原成对人友好的自然语言。简单的 NLG 方法可以是事先设定好的回复模板，复杂的可以使用深度学习生成模型，如“Semantically Conditioned LSTM”通过在 LSTM 中加入对话动作 cell 辅助答复生成。

任务型对话机器人最权威的研究者是剑桥大学的 Steve Young 教授，强烈推荐他的教程“Statistical Spoken Dialogue Systems”。他的诸多博士生针对上面各个流程都做了很细致的研究，想了解细节的话可以参考他们的博士论文。相关课程可参考 Milica Gašić 的“Speech and Language Technology”。

除了把整个问题分解成上面几个流程分别优化，目前很多学者也在探索使用端到端技术整体解决这个问题，代表工作有 Tsung-Hsien Wen 等人的“A Network-based End-to-End Trainable Task-Oriented Dialogue System”和 Xiujun Li 等人的“End-to-End Task-Completion Neural Dialogue Systems”，非常值得学习。

2.4 闲聊型机器人

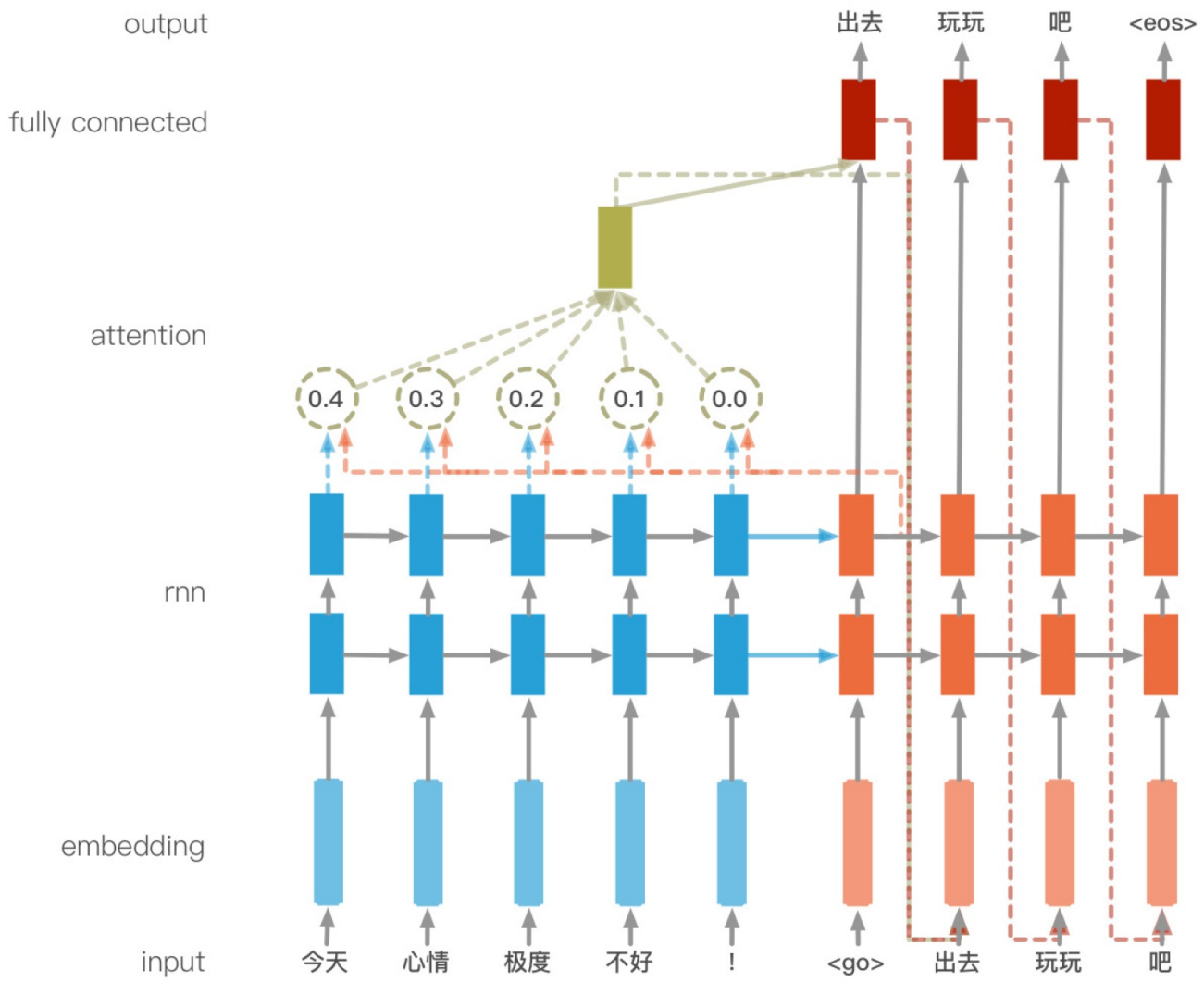

图4 Chitchat—Bot的seq2seq模型框架

真实应用中，用户与系统交互的过程中不免会涉及到闲聊成分。闲聊功能可以让对话机器人更有情感和温度。闲聊机器人（ Chitchat-Bot ）通常使用机器翻译中的深度学习 seq2seq 框架来产生答复，如图4。

与机器翻译不同的是，对话中用户本次 query 提供的信息通常不足以产生合理的答复，对话的历史背景信息同样很重要。例如图4中的 query：“今天心情极度不好!”，用户可能是因为前几天出游累的腰酸背痛才心情不好的，这时答复“出去玩玩吧”就不合情理。研究发现，标准的 seq2seq+attention 模型还容易产生安全而无用的答复，如“我不知道”，“好的”。

为了让产生的答复更多样化、更有信息量，很多学者做了诸多探索。Jiwei Li 等人的论文“Deep Reinforcement Learning for Dialogue Generation”就建议在训练时考虑让答复引入新信息，保证语义连贯性等因素。Iulian V. Serban 等人的文“Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models”在产生答复时不只使用用户当前 query 的信息，还利用层级 RNN 把之前对话的背景信息也加入进来。Jun Yin 等人的论文“Neural GenerativeQuestion Answering”在产生答复时融合外部的知识库信息。

上面的各种机器人都是为解决某类特定问题而被提出的，我们前面也分开介绍了各个机器人的主要组件。但这其中的不少组件在多种机器人里都是存在的。例如知识图谱在检索型、任务型和闲聊型机器人里也都会被使用。

真实应用中通常会包含多个不同类型的机器人，它们协同合作，解答用户不同类型的问题。我们把协调不同机器人工作的机器人称之为路由机器人（ Route-Bot ）。路由机器人根据历史背景和当前 query，决定把问题发送给哪些机器人，以及最终使用哪些机器人的答复作为提供给用户的最终答复。图5为框架图。

2.5 对话机器人现状

对话机器人历史悠久，从1966年 MIT 的精神治疗师机器人 ELIZA 到现在已有半个世纪。但现代意义的机器人其实还很年轻。检索型单轮对话机器人得益于搜索引擎的商业成功和信息检索的快速发展，目前技术上已经比较成熟。最近学术界和工业界也积极探索深度学习技术如 Word2vec、CNN 和 RNN 等在检索型机器人中的使用，进一步提升了系统精度。虽然技术上较为成熟，但在实际应用中检索型机器人还存在不少其他问题。例如，很多企业历史上积累了大量非结构化数据，但这些数据并不能直接输进检索型机器人，而是需要事先通过人工整理。即便有些企业存在一些回答对的数据可以直接输入检索型机器人，但数量往往只有几十到几百条，非常少。可用数据的质量和数量限制了检索型机器人的精度和在工业界的广泛使用。

相较于检索型机器人，知识图谱型机器人更加年轻。大多数知识图谱型机器人还只能回答简单推理的事实类问题。这其中的一个原因是构建准确度高且覆盖面广的知识图谱极其困难，需要投入大量的人力处理数据。深度学习模型如 Memory Networks 等的引入可以绕过或解决这个难关吗？

任务型多轮对话机器人只有十来年的发展历史，目前已能较好地解决确定性高的多轮任务。但

当前任务型机器人能正常工作的场景往往过于理想化，用户说的话大部分情形下都无法精确表达成 act-slot-value 三元组，所以在这个基础上构建的后续流程就变得很脆弱。很多学者提出了各种端到端的研究方案，试图提升任务型机器人的使用鲁棒性。但这些方案基本都需要利用海量的历史对话数据进行训练，而且效果也并未在真实复杂场景中得到过验证。

开域闲聊型机器人是目前学术界的宠儿，可能是因为可改进的地方实在太多吧。纯粹的生成式模型在答复格式比较确定的应用中效果已经不错，可以应用于生产环境；但在答复格式非常灵活的情况下，它生成的答复连通顺性都未必能保证，更不用说结果的合理性。生成模型的另一个问题是它的生成结果可控性较低，效果优化也并不容易。但这方面的学术进展非常快速，很多学者已经在探索深度增强学习、GAN 等新算法框架在其上的使用效果。

虽然目前对话机器人能解决的问题非常有限，短期内不可能替代人完成较复杂的工作。但这并不意味着我们无法在生成环境中使用对话机器人。寻找到适宜的使用场景，对话机器人仍能大幅提升商业效率。截止到目前，爱因互动已经成功把对话机器人应用于智能投顾、保险、理财等销售转化场景，也在电商产品的对话式发现和推荐中验证了对话机器人的作用。 如果一个对话机器人与真人能顺利沟通且不被真人发现自己是机器人，那么就说这个机器人通过了图灵测试。当然目前的对话机器人技术离这个目标还很远，但我们正在逐渐接近这个目标。随着语音识别，NLP 等技术的不断发展，随着万物互联时代的到来，对话机器人的舞台将会越来越大。

文/吴金龙