1、A Network-based End-to-End Trainable Task-oriented Dialogue System

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

当前构建一个诸如宾馆预订或技术支持服务的 task-oriented 的对话系统很难，主要是因为难以获取训练数据。现有两种方式解决问题：

* 将这个问题看做是 partially observable Markov Decision Process (POMDP)，利用强化学习在线与真实用户交互。但是语言理解和语言生成模块仍然需要语料去训练。而且为了让 RL 能运作起来，state 和 action space 必须小心设计，这就限制了模型的表达能力。同时 rewad function 很难设计，运行时也难以衡量
* 利用 seq2seq 来做，这又需要大量语料训练。同时，这类模型无法做到与数据库交互以及整合其他有用的信息，从而生成实用的相应。

本文提出了平衡两种方法的策略。

2、How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

3、End-to-End Reinforcement Learning of Dialogue Agents for Information Access

2016年卡耐基梅隆大学研究团队利用深度强化学习进行对话状态追踪和管理

4、Zhao T, Eskenazi M. Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning [J]. arXiv preprint arXiv:1606.02560, 2016.

2016年麻省理工大学研究团队提出层次化DQN模型，其代码采用Keras实现并开源[ code ]，该工作发表在NIPS2016上。

5、Kulkarni T D, Narasimhan K R, Saeedi A, et al. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation [J]. arXiv preprint arXiv:1604.06057, 2016.

6、BBQ-Networks: Efficient Exploration in Deep Reinforcement Learning for Task-Oriented Dialogue Systems

AAAI2018 录用文章，将深度强化学习用于对话系统。BBQ network 这个名字很有意思，工作来自微软研究院和 CMU。

提出了一种新的算法，可以显著提升对话系统中深度 Q 学习智能体的探索效率。我们的智能体通过汤普森采样（Thompson sampling）进行探索，可以从 Bayes-by-Backprop 神经网络中抽取蒙特卡洛样本。我们的算法的学习速度比 ε-greedy、波尔兹曼、bootstrapping 和基于内在奖励（intrinsic-reward）的方法等常用的探索策略快得多。此外，我们还表明：当 Q 学习可能失败时，只需将少数几个成功 episode 的经历叠加到重放缓冲（replay buffer）之上，就能使该 Q 学习可行。

7、Deep Reinforcement Learning with Double Q-Learning

8、Deep Attention Recurrent Q-Network

9、SimpleDS: A Simple Deep Reinforcement Learning Dialogue System

10、Deep Reinforcement Learning with a Natural Language Action Space

11、**Integrating User and Agent Models: A Deep Task-Oriented Dialogue System**

**12、A Deep Reinforcement Learning Chatbot**

蒙特利尔算法研究实验室（MILA）为参与亚马逊 Alexa 大奖赛而开发的深度强化学习聊天机器人。

MILABOT 能够与人类就流行的闲聊话题进行语音和文本交流。该系统包括一系列自然语言生成和检索模型，如模板模型、词袋模型、序列到序列神经网络和隐变量神经网络模型。

通过将强化学习应用到众包数据和真实用户互动中进行训练，该系统学习从自身包含的一系列模型中选择合适的模型作为响应。

真实用户使用 A/B 测试对该系统进行评估，其性能大大优于竞争系统。由于其机器学习架构，该系统的性能在额外数据的帮助下还有可能继续提升。