# 面向任务的对话系统异步更新强化学习框架

*Sai Zhang，Yuwei Hu，Xiaojie Wang** 和 *Caixia Yuan*

中国北京，北京邮电大学{zs,hyw724,xjwang,yuancx}
@bupt.edu.cn

arXiv:2305.02718v1 [cs.CL] 4 May 2023

**摘要**

许多研究已将强化学习应用于对话系统的训练。以往的方法将对话系统划分为多个模块，包括 DST（对话状态跟踪）和 DP（对话策略），并同时训练这些模块。然而，在训练过程中，不同模块会相互影响。DST 产生的错误可能会误导对话策略，而系统动作也会给 DST 模块带来额外困难。为缓解这一问题，我们提出**异**步**更**新**强**化**学**习框架（AURL），在协作设置下异步更新对话状态跟踪模块和对话策略模块。此外，我们采用课程学习来解决强化学习采样过程中数据分布不均衡的问题，并引入多个用户模型以提高对话多样性。在公开的 SSD-PHONE 数据集上的实验结果表明，我们的方法取得了令人瞩目的成绩，对话成功率提高了 31。37%。代码已公开发布于 https://github.com/shunjiu/AURL。

索引词— 面向任务的对话系统、多智能体强化学习、课程学习、用户模拟器

## 1. 引言

面向任务的对话系统广泛应用于客户服务，例如自动订票。对话系统通常由四个模块组成：自然语言理解（NLU）、对话状态跟踪（DST）、对话策略（DP）和自然语言生成（NLG）[1]。DST 负责从对话开始到当前轮次维护对话状态，通常采用监督学习（SL）进行训练。DP 决定系统动作以引导对话方向，可通过带标签的数据进行监督学习（SL）[2]，也可通过强化学习（RL）[3, 4] 进行训练，此时用户模拟器作为交互环境的一部分。

在以往的工作中，使用强化学习训练对话策略时，通常会应用基于规则的对话状态跟踪 [4, 5]，以忽略对话状态跟踪中的错误。然而，在真实场景中，对话状态跟踪模块并不稳定。对话策略应在对话状态跟踪错误的影响下引导对话成功。 [6, 7] 使用监督学习和强化学习损失训练这些模块from

*通讯作者。

| 级别 | | 对话 |
|---|---|---|
| Easy | 上一轮系统状态 | [159] |
| | 上一条系统话语 | 然后呢?(And then?) /需要更多/ |
| | 用户话语 | 4307 /信息规范/ |
| | 生成的信念状态 | [159, 4307] |
| | 预言信念状态 | [159, 4307] |
| Hard | 上一轮系统状态 | [159, 4307] |
| | 上一条系统话语 | 那应该是多少(应该是多少？) /请求更正/ |
| | 用户话语 | 我记错了，是807，307错了 /更新子项/ (我记错了，是807，307是错的。) |
| | 生成的信念状态 | [159, 807] |
| | 预言信念状态 | [159, 4807] |

**图 1**：简单和困难对话示例，不同的用户动作会给系统的对话状态跟踪模块带来不同挑战。用户动作 [inform norm] 对系统而言较简单，而 [update sub] 则较困难。对话状态跟踪模块产生的错误可能会误导系统对话策略模块的决策。用户动作与系统对话策略模块所决定的系统动作相关。

对话策略。然而，在同时训练时，这些模块会相互影响。如果 DST 跟踪到错误的槽位值，系统可能会在正确的策略下收集到错误的槽位，从而导致策略学习出现偏差。与此同时，假设对话策略模块经常选择用户容易作出回应的动作，那么 DST 模块就无法得到充分训练，并可能失去理解困难用户动作的能力。例如，图 1 展示了不同的用户动作会给 DST 模块带来不同的挑战，从而在训练过程中引入偏差。此外，对话策略通常通过与预定义的用户模拟器 [5], 进行交互来训练，尽管该模拟器具有随机性，但其行为仍较为单一。正如 [6], 中所观察到的，不同用户会带来多样化的对话状态，因此能够使对话策略得到充分训练。

为克服这些问题，我们首先提出一种用于对话系统的新型异步更新强化学习框架，其中对话状态跟踪和对话策略模块均可协同训练，但各自以不同的更新频率异步更新。随后，我们基于 AURL 框架构建了多用户强化学习（MURL），使用多个用户模型与一个对话系统交互，从而获得更多样化的对话策略。此外，我们采用课程

学习 [8]，通过在训练过程中逐步提高所用数据样本的复杂度，促进对话状态跟踪的学习。

为更好地验证我们的方法，我们在 SSD-PHONE 数据集上开展实验 [9], 。这是一个具有多样性现象的大动作空间对话数据集。我们构建了一个分层神经网络对话系统和一个用户模型。结果表明，使用所提出的框架训练的系统在在线测试中达到了新的最先进水平。总而言之，本文的贡献如下。

- 我们提出了一种新颖的对话系统异步更新强化学习框架，在协作设置下异步训练对话策略和对话状态跟踪模块，并应用课程学习解决训练过程中的偏差。

- 我们提出了一种新颖的强化学习框架，用于使用多个用户模型训练一个对话代理，从而提高对话多样性，并充分促进对话策略学习。

- 我们在真实世界的 SSD-PHONE 数据集上进行了实验。结果表明，我们的方法优于多种强基线方法。值得注意的是，与当前最先进水平相比，我们的方法使对话成功率提高了 31.37%。

## 2. 方法

### 2.1. 异步更新强化学习框架

#### *2.1.1.* 框架

图 2 展示了我们框架的总体架构和信息流，该框架由一个系统代理和 $N$ 个用户代理组成。系统与用户通过脚本语言相互通信。完成一轮对话后，每个模块的输入、输出、标签或奖励分别存储在回放缓冲区 $BF_{user\ DP}$，$BF_{user\ NLU}$，$BF_{sys\ DP}$和 $BF_{sys\ DST}$中。我们使用经验回放[10]异步更新每个模块。

**系统对话状态跟踪** 生成 $\phi$，用于更新当前轮次的对话信念状态。受 TRADE [11], 启发，我们使用两个编码器分别对对话历史 $[U^s_{t-1};\ U^u_t]$和上一轮系统信念状态 $BS^s_{t-1}$ 进行编码，其中 $U^s_{t-1}$ 是第 $t-1$ 轮的系统响应，$U^u_t$是当前轮次的用户话语 $t$。我们使用状态生成器 [12] 生成当前轮次的 $\widehat{\ }$ 信念状态 $BS^s_t$。用户动作 $a^u_t$ 由一个多层感知器生成，其输入为两个编码器的隐状态。系统对话状态跟踪的公式如下：

$$(BS^s_t, \widehat{a^u_t}) = \phi(BS^s_{t-1}, [U^s_{t-1}; U^u_t]). \tag{1}$$

**系统对话策略**生成 $\pi$，用于根据对话状态决定当前轮次的系统动作 $a^s_t$ 和系统槽位 $s^s_t$。

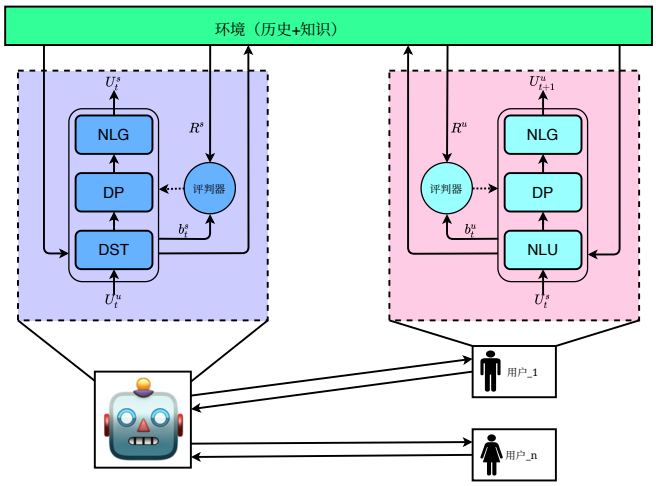


**图 2**：基于多智能体学习的 AURL 框架。系统与用户通过书面语言进行交互。系统的对话状态跟踪模块和对话策略模块异步更新。

对话轮次 $t$ 的对话状态由以下部分拼接而成：(1) 上一轮次的系统动作 $a^s_{t-1}$，(2) 当前轮次的信念状态 $BS^s_t$，(3) 对话历史编码器的隐藏状态 $H^{ctx}_t$，(4) 来自数据库的查询结果 $q_t$。系统对话策略表示为：

$$(a^s_t, s^s_t) = \pi(a^s_{t-1}; BS^s_t; H^{ctx}_t; q_t). \tag{2}$$

**用户自然语言理解** 生成 $\eta$，用于根据系统话语 $U^s_t$ 理解系统动作 $a^s_t$ 和系统槽位 $s^s_t$ 以及用户目标值 $G$。用户自然语言理解的形式化定义如下：

$$(\widehat{a^s_t}, \widehat{s^s_t}) = \eta(G; U^s_t). \tag{3}$$

**用户对话策略** 生成 $\mu$，用于确定用户动作 $a^u_{t+1}$ 和用户槽位 $s^u_{t+1}$，以便与系统代理进行交互。随后，状态向量 $BS^u_{t+1}$ 记录每个槽位或槽位的每个部分是否已提供或需要更新。每个槽位采用 0-2 的李克特量表，分别表示未提供、已提供 和需要更新。为模拟自动语音识别（ASR）带来的错误，我们会随机将槽位替换为一个相似的槽位。用户策略模块的输入是以下内容的拼接：(1) 上一轮的用户动作 $a^u_t$、(2) 系统动作 $a^s_t$、(3) 系统槽位 $s^s_t$ 以及 (4) 上一轮的用户状态向量$BS^u_t$。用户策略的形式化定义如下：

$$(a^u_{t+1}, s^u_{t+1}) = \mu(a^u_t; \widehat{a^s_t}; \widehat{s^s_t}; BS^u_t). \tag{4}$$

为减轻对话状态跟踪与对话策略之间的错误纠缠，我们以不同的更新频率异步训练这两个模块。我们通过经验回放训练这两个模块：当经验池装满时，从中采样训练样本。对话策略使用较小经验池中的样本进行更频繁的更新，而对话状态跟踪则使用较大经验池中的样本缓慢更新。这样，对话策略便能在相对稳定的环境中快速优化其策略，而对话状态跟踪则能使用最新的对话策略模块进行学习。

#### *2.1.2.* 课程学习

在强化学习过程中，训练样本从经验池中随机采样。为了模拟人类的学习方式，我们引入课程学习来安排训练过程。在离线对话数据集上完成预训练后，使用测试数据集评估对话状态跟踪模型，并得出用户行为理解的联合准确率。随后，将联合准确率用作难度衡量指标 [13]，以此将 $BF_{sys\ DST}$中的数据划分为简单、中等 和困难 等级。

首先，我们依次使用简单、中等 和困难 级别的数据训练对话状态跟踪模块。其次，由于简单 数据占全部数据的 75%，且模型已在这些数据上充分学习，因此使用中等 和困难级别的数据。第三，仅使用困难 数据。最后，依次利用所有级别的数据进行复习。

#### *2.1.3. MURL*

使用多个在对话语料库上预训练的用户模型来训练一个对话系统。系统完成一轮对话后，会依次与各个用户进行交互。由于随机数的影响，这些彼此独立训练的用户在强化学习训练过程中会表现出不同的个性。对于相同的系统话语，不同用户可能会采取不同的动作进行回应，从而增加对话的多样性，促进系统的对话策略学习。

### 2.2. 奖励

奖励对于强化学习至关重要，可用于引导策略学习。系统和用户承担的角色不同。对于面向任务的对话任务，系统应以较少的轮次成功完成对话。c 在我们的设置中，系统与用户通过协作交流来完成对话。以下展示了针对各角色的奖励设置。

对于系统奖励 $R^s$，它由以下部分组成：(1) 对话结束时的对话成功奖励和对话失败惩罚；(2) 每轮较小的对话长度惩罚；(3) 行为与槽位不匹配惩罚，以避免系统确认空槽位值，等等；(4) 根据用户动作设置的少量不恰当系统动作惩罚。

与系统不同，用户策略应具有多样性，并适合促进系统策略学习。用户奖励 $R^u$ 与系统奖励相似。我们去除了针对用户的对话轮次长度惩罚，因为我们无法限制用户快速结束对话。

### 2.3. 优化

算法 1 展示了在单个用户设置下完整的 AURL 算法。对于系统对话状态跟踪模块和用户自然语言理解模块，我们可以在两者交互时获得标签输出。因此，我们使用交叉熵损失来更新这两个模块。

**算法 1：** 包含一个用户的 AURL 框架。

**输入：** 对话语料库 $D$；系统模型和用户模型；系统对话状态跟踪缓冲区大小 $n$

**输出：** 训练后的系统模型。

随机初始化权重 $\phi, \pi, \eta, \mu, V^s, V^u$；
使用监督学习在对话语料库 $D$ 上预训练 $\phi, \pi, \eta, \mu$。
**对于每个**训练轮次 **执行**
  **对于每个** 步骤**执行**
    初始化用户目标值 $G$、用户状态和系统状态。
    系统在第一轮给出话语 $U_0^s$。
    **重复**
      用户理解系统话语 $U^s$，使用 $\eta$、$\mu$ 对动作和槽位进行采样，并给出响应 $U^u$。
      系统更新其对话状态 $BS^s$
      根据用户响应，使用 $\phi$，然后使用 $\pi$采样动作和槽位，给出响应 $U^s$。
      根据 $BS^s$获取终止信号 T，$G$和对话长度。
      观察奖励 $R^s$ 和 $R^u$。
      四个回放缓冲区记录输入、输出、每个模块的标签或奖励。
    **直到**根据 *T*; 对话结束
  **end**
  更新两个评论家网络、两个对话策略模块和用户自然语言理解模块。清空三个缓冲区·
  **如果** $|BF_{sys\ DST}|$ 等于 $n$，**那么**
    使用课程学习更新系统对话状态跟踪模块学习。明确 $buffer_{system\ DST}$
  **end**
**结束**

优势演员-评判器（A2C）算法用于优化两个策略模块。对于每个角色，使用评判器网络 $V$ 评估状态价值。评判器网络旨在最小化以下损失函数：

$$L_V^s = (R^s + \gamma V^s(b_{t+1}^s) - V^s(b_t^s))^2, \tag{5}$$

$$L_V^u = (R^u + \gamma V^u(b_{t+1}^u) - V^u(b_t^u))^2, \tag{6}$$

其中 $b_t^s = [a_{t-1}^s;\ BS_t^s;\ H_t^{history};\ q_t]$, $b_t^u = [a_t^u;\ a_t^s;\ s_t^s;\ BS_t^u]$。

演员网络（策略）的目标是最大化回报。优势函数 $A(a, s; b_t) = R + \gamma V(b_{t+1}) - V(b_t)$ 用于评估所选动作和槽位是否更优。策略模块的损失函数如下：

$$L_\pi^s = A(a^s; s^s, b^s)(\log_\pi(a^s|b^s) + \log_\pi(s^s|b^s)), \tag{7}$$

$$L_\mu^u = A(a^u; s^u, b^u)(\log_\mu(a^u|b^u) + \log_\mu(s^u|b^u)). \tag{8}$$

此外，我们使用不同大小的回放缓冲区，异步更新系统的对话状态跟踪及其他模块。

## 3. 实验

### 3.1. 实验设置

**数据集**。 SSD-PHONE [9] 是一个真实世界的面向任务的对话语料库，包含 30 种动作、11,000 个对话会话、3,135 条不同的对话路径以及丰富的多样性现象。该语料库还提供了大量标注信息。为了验证强化学习在更具挑战性的场景中的能力，我们根据多样性现象进一步将对话动作的数量扩展到 41 种，包括 16 种系统动作和 25 种用户动作。

**评估指标**。 我们通过与 [9] 提供的基于有限状态自动机的用户模拟器进行在线交互来评估模型性能。每次与 1,000 个对话交互后，共交互三次，我们计算以下指标。**对话成功率** 是主要指标。当且仅当系统在限定轮次内收集到的槽值与用户目标值相等时，对话才算成功。**平均轮数** 表示成功对话的平均轮数。**平均奖励** 是每个对话的系统奖励的平均值。**对话状态跟踪准确率**表示每轮是否正确收集了槽值。**平均时间** 用于衡量与用户交互时的平均响应时间。

**实现细节**。 回复缓冲区的大小分别为 6、6、6、$3k$，对应 $BF_{user\ DP}$、 $BF_{user\ NLU}$、$BF_{sys\ DP}$ 和 $BF_{sys\ DST}$。在强化学习训练期间，训练 $100k$ 个轮次，每个轮次包含 6 个对话。在奖励设计方面，如果对话成功，则系统和用户的奖励均设为 2.0；否则，两者均设为 $-1.0$。系统的其他类型惩罚项设为 $-0.05$，用户的所有其他类型惩罚项均设为 $-0.02$。 $\gamma$ 设为 0.99。

### 3.2. 基线方法

除了 [9], 中的基线方法外，我们还将 AURL 与不同方法进行比较。**DAMD**[2] 使用监督学习训练对话策略，因为一个对话状态可能对应多个系统动作。**监督学习** 首先是在 SSD-PHONE 数据集上预训练的系统模型。以下所有基线方法均使用预训练的系统模型和用户模型。我们使用 REINFORCE [14], 、A2C、异步优势演员-评论家[15] 和近端策略优化 [16] 算法进行实验，最终采用 A2C 作为学习算法，因为 A2C 的学习曲线更加稳定。**RL-fixed DST** 是传统的强化学习设置，仅更新系统策略模块和用户策略模块。**RL-train DST**表示同时训练所有模块。

### 3.3. 结果与分析

各模型的在线评估结果汇总于表1中。在监督学习设置下，所提出的模型表现最佳，相较于基于 GPT2 [17] 的模型（SimpleTOD [18]）提升了 11.40%。此外，我们的模型比其他基线方法更加轻量，且具有更短的每轮响应时间，这使得在真实对话平台上部署时与用户交互能够更快地给出响应。

| 模型 | 对话 succ | Avg turn | Avg 奖励 | DST acc | AVG time |
|---|---|---|---|---|---|
| TRADE* | 30.45 | 9.77 | - | - | 111 |
| DAMD | 46.40 | 6.52 | - | - | 489 |
| UBAR | 57.70 | 11.39 | - | - | 376 |
| SimpleTOD | 63.20 | 8.18 | - | - | - |
| SL | **75.73** | 7.35 | 1.19 | 71.12 | **28** |
| RL-fixed DST | 77.23 | 6.93 | 1.28 | 66.78 | - |
| RL-train DST* | 71.00 | **5.94** | 1.11 | 75.58 | - |
| AURL | 84.13 | 7.36 | 1.47 | 70.92 | - |
| AURL-1v2 | **94.57** | 8.32 | **1.82** | **81.11** | - |

**表1**：不同模型与基于有限状态自动机的用户模拟器交互的结果。AURL-1v2 表示使用两个用户模型训练系统模型。其他强化学习设置下仅使用一个用户模型。 $\star$ 为消融研究。

在 RL 设置下，RL-fix DST 与 RL-train DST 的对比表明，同时训练方法存在偏差。使用一个用户模型训练对话系统时，AURL 的对话成功率高于 SL，提升了 8.40%。借助异步更新和课程学习，对话策略模块甚至能够在对话状态跟踪出错时做出正确决策。

使用两个用户模型训练一个对话系统时，对话成功率从 84.13% 提升至 94.57%。此外，对话状态跟踪模块的准确率达到 81.11%。更多用户模型带来更多不同的对话路径，就像现实中的各种人类一样，尽管它们使用相同的预训练参数进行初始化。与仅使用一个用户模型相比，对话系统得到了更充分的训练。

## 4. 结论

本文提出了一种用于面向任务的对话系统中对话状态跟踪和对话策略模块的异步更新强化学习框架。我们在异步更新框架中开展多智能体强化学习，以训练这两个模型。借助课程学习和多个用户模型训练，我们的方法取得了新的最先进水平，在在线测试中比原始的 GPT2 模型提升了 31.37%。未来，我们将在 MURL 设置下开展更多工作，尤其是引入更多用户模型来训练系统代理。

## 5. 致谢

我们感谢匿名审稿人提出的深刻意见。本研究得到中国国家自然科学基金重大研究计划（项目编号：92067202）的支持。

## 6. 参考文献


[1] B. Liu 和 I. Lane，“面向任务的对话端到端学习”，载于*2018*年北美计算语言学协会分会年会论文集：学生研究研讨会。美国路易斯安那州新奥尔良：计算语言学协会，2018年6月，第67–73页。

[2] Y. Zhang、Z. Ou 和 Z. Yu，“考虑同一上下文下多个适当响应的任务型对话系统”，载于*AAAI* 人工智能会议论文集，第 34 卷第 05 期，2020 年，第 9604–9611 页。

[3] R. Takanobu、H. Zhu 和 M. Huang，“引导式对话策略学习：面向多领域任务型对话的奖励估计”，载于*2019* 年自然语言处理经验方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集。中国香港：计算语言学协会，2019 年 11 月，第 100–110 页。

[4] R. Takanobu、R.Liang 和 M. Huang，“采用角色感知奖励分解的多智能体任务型对话策略学习”，载于计算语言学协会第 *58* 届年会论文集，2020 年，第 625–638 页。

[5]Z. Tang、H. Kulkarni 和 G. H. Yang，“通过间歇式短扩展集成实现高质量对话多样化”，载于计算语言学协会会议论文集：*ACL-IJCNLP 2021* 论文集。线上：计算语言学协会，2021年8月，第1861–1872页。

[6] S. Lei、X. Wang 和 C. Yuan，“结合对话知识的协作式多智能体强化学习对话管理”，应用科学，第10卷，第8期，第2740页，2020年。

[7] H. Liu、C. Yuan、X. Wang、Y. Yang、H. Jiang 和 Z. Wang，“对话、聚焦与猜测——面向多文档驱动的对话”，载于*AAAI* 人工智能会议论文集，第35卷，第15期，2021年，第13 380–13387页。

[8] Y. Bengio、J. Louradour、R. Collobert 和 J. Weston，《课程学习》，载于第*26*届国际机器学习年度会议论文集，2009年，第41–48页。

[9] S. Zhang、Y. Hu、Y. Wu、J. Wu、Y. Li、J. Sun、C. Yuan 和 X. Wang，《一个槽位并非由一次话语构建：带子槽位的口语对话》，载于计算语言学协会：*ACL* 论文集 *2022*年。爱尔兰都柏林：计算语言学协会，2022年5月，第309–321页。[10]T. Schaul、J. Quan、I. Antonoglou 和 D. Silver，《优先经验回放》，*arXiv* 预印本 *arXiv:1511.05952*，2015年。

[11]C.-S. Wu、A. Madotto、E. Hosseini-Asl、C. Xiong、R. Socher 和 P. Fung，《面向任务的对话系统的可迁移多领域状态生成器》，载于第*57*届计算语言学协会年会论文集。意大利佛罗伦萨：计算语言学协会，2019年7月，第808–819页。

[12] M. Heck、C. van Niekerk、N. Lubis、C. Geishauser、H.-C. Lin、M. Moresi 和 M. Gasic，《Trippy：一种用于与值无关的神经对话状态跟踪的三重复制策略》，载于第*21*届话语与对话特别兴趣小组年会论文集。首届线上会议：计算语言学协会，2020年7月，第35–44页。

[13] X.Wang、Y. Chen 和 W. Zhu，《课程学习综述》，*IEEE*模式分析与机器智能汇刊，2020。

[14] R. J. Williams，《用于连接主义强化学习的简单统计梯度跟随算法》，强化学习，第 5–32 页，1992。

[15] V. Mnih、A. P. Badia、M. Mirza、A. Graves、T. Harley、T. P. Lillicrap、D. Silver 和 K. Kavukcuoglu，《深度强化学习的异步方法》，载于第 *33* 届国际机器学习会议论文集——国际机器学习会议第 *48* 卷，ICML’16 系列。JMLR.org，2016，第 1928–1937 页。

[16] J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov，《近端策略优化算法》，*arXiv* 预印本 *arXiv:1707.06347*，2017。

[17] A. Radford、J. Wu、R. Child、D. Luan、D. Amodei、I. Sutskever等，《语言模型是无监督多任务学习器》。

[18] E. Hosseini-Asl、B. McCann、C.-S. Wu、S. Yavuz 和 R. Socher，《一种用于面向任务的对话的简单语言模型》，载于神经信息处理系统进展，H. Larochelle、M. Ranzato、R. Hadsell、M. F. Balcan 和 H. Lin 编，第 33 卷。Curran Associates, Inc.，2020 年，第 20 179–20191 页。