深度加固学习Chatbot

Iulian V. Serban，Chinnadhurai Sankar，Mathieu Germain，Saizheng Zhang，Zhouhan Lin，

Sandeep Subramanian，Taesup Kim，Michael Pieper，Sarath Chandar，Nan Rosemary Ke，

Sai Mudumba，Alexandre de Brebisson Jose MR Sotelo，Dendi Suhubdy，

Vincent Michalski，Alexandre Nguyen，Joelle Pineau和Yoshua Bengio

蒙特利尔学习算法研究所，蒙特利尔，魁北克，加拿大

抽象

我们提出了MILABOT：由蒙特利尔学习算法研究所（MILA）为亚马逊Alexa奖比赛开发的深度加强学习chatbot。MILABOT能够通过言语和文字与人类交流流行的小谈话话题。该系统由自然语言生成和检索模型组成，包括基于模板的模型，袋字模型，序列到神经网络和潜变量神经网络模型。通过将强化学习应用于众包数据和现实世界的用户交互，系统已经被训练以从其集合中的模型中选择适当的响应。该系统已经通过对真实用户的A / B测试进行评估，其性能明显优于竞争系统。由于其机器学习架构，

1        简介

对话系统和会话代理（包括聊天室，个人助理和语音控制界面）在现代社会中变得普遍存在。这些例子包括移动设备上的个人助理，电话线上的技术支持帮助，以及从时尚服装和化妆品销售任何东西的在线漫游器到法律咨询和自助治疗。然而，构建智能会话代理仍然是人工智能研究中未解决的主要问题。

2016年，亚马逊公司提出了一个国际大学竞赛，目的是建立一个社交机构：一个口语会话代理，能够与人类一致地和人性地交流，如娱乐，时尚，政治，体育和技术等热门话题。社交工作者通过亚马逊的回声设备（Stone＆Soper 2014）进行自然语言演讲。本文介绍了我们在蒙特利尔大学开发的模型，实验和最终系统（MILABOT）。1我们参与的主要动机是帮助推进人工智能研究。为此，比赛为真正用户（也称为*野外机器学习）*培训和测试最先进的机器学习算法提供了一个特殊的机会）在一个相对无约束的环境。对真实用户进行实验的能力在人工智能社区是独一无二的，绝大多数工作包括对固定数据集（例如标记的数据集）和软件模拟（例如游戏引擎）的实验。此外，亚马逊提供的计算资源，技术支持和财务支持有助于扩大我们的系统，并测试最先进的机器学习方法的极限。别的不说，这种支持使我们能够以众包200 *，* 000 在亚马逊的Mechanical Turk标签，并保持了32个专用的Tesla K80 GPU的运行我们的直播系统。

1

我们的团队被称为MILA团队，MILA代表蒙特利尔学习算法学院。

Alexa奖第一辑（Alexa奖2017）。

我们的社会机构基于利用深入学习和强化学习的大型综合系统。我们为自然语言检索和生成开发了一套新的深度学习模型

- 包括循环神经网络，序列到序列模型和潜在变量模型，并在竞争的背景下进行评估。这些模型被组合成一个集合，它产生一组候选对话响应。此外，我们应用强化学习，包括价值功能和策略梯度方法，以训练系统，从其合奏中的模型中选择适当的反应。特别是，我们提出了一种基于马尔可夫决策过程估计的新型强化学习过程。对众包数据和现实世界用户之间记录的交互以及系统的初步版本进行培训。受过训练的系统在真实用户的A / B测试实验中产生了实质性的改进。

在比赛半决赛中，我们表现最好的系统达到了3 分的平均用户分数*。*15 级，1 - 5级，手工制作国家和规则数量最少，不进行*非对话活动*（如玩游戏或参加测验）。[[1]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn1" \o ")这个最佳系统的表现与半决赛中的顶级系统相当。[[2]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn2" \o ")此外，同一系统平均高14 *。*5 - 16 *。*0每个对话转动 用户和系统之间的前后交流的这种改进表明，我们的系统可能是竞争中所有系统中最具互动性和吸引力的系统。最后，由于几乎所有的系统组件都可以学习，所以系统必将随着附加数据而改进。

2        系统概述

早期的对话系统工作（Weizenbaum 1966，Colby 1981，Aust et al。1995，McGlashan et al。1992，Simpson＆Eraser 1993）主要基于人类专家手工制作的国家和规则。现代对话系统通常遵循混合架构，将手工制作的状态和规则与统计机器学习算法相结合（Suendermann-Oeft et al。，2015，Jurcícek et al。，2014，Bohus et al。2007，Williams 2011）。然而，由于人类语言的复杂性，可能无法枚举在开放域名流行主题上构建能够与人交谈的社会机构所需的状态和规则。与这种基于规则的系统相比，我们的核心方法完全基于统计机器学习。我们相信，这是人为智能对话代理人最可信的途径。我们提出的系统架构旨在尽可能少地做出关于理解和产生自然人类语言过程的假设。因此，该系统仅使用少量的手工制作状态和规则。然而，每个系统组件都被设计为使用机器学习算法进行优化（训练）。这些系统组件将首先在大量数据集上独立进行培训，然后联合实际的用户交互。这样，系统将学习隐含进行开放域对话的所有相关状态和规则。给出足够的例子，这样的系统应该优于基于手工制作的状态和规则的系统。此外，系统将继续随着附加数据的不断改进。

我们的系统架构受到基于组合的机器学习系统的成功的启发。这些系统由许多独立的子模型组合在一起。这样的系统系统的例子包括Netflix奖的获胜者（Koren等人，2009），利用数百种机器学习模型来预测用户电影偏好，以及IBM Watson（Ferrucci等人2010），第一个获得机器学习系统测验游戏Jeopardy！最近，谷歌观察到大量改进，构建了一个基于整体的神经机器翻译系统（Wu et al。2016）。

我们的系统由一个响应模型组成。响应模型以输入对话形式输出自然语言文本中的响应。另外，响应模型还可以输出一个或几个标量值，表明它们的内部置信度。如后面将要解释的那样，响应模型被设计成使用各种策略在不同的主题集上产生响应。作为输入，整个系统期望被给予对话历史（即迄今为止记录在对话中的所有话语，包括当前用户话语）和自动语音识别系统的置信度值（ASR信号）。要生成响应，系统如下

图1：对话管理员控制流程。

一个三步骤的程序。首先，它使用所有响应模型来生成一组候选响应。第二，如果在候选响应集合中存在*优先级*响应（即优先于其他响应的响应），则该响应将由系统返回。[[3]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn3" \o ")例如，对于问题*“你叫什么名字？”*，*“我是Alexa奖Socialbot”*的回应是一个优先的回应。第三，如果没有*优先级*响应，则由*模型选择策略*选择响应。例如，*模型选择政策*可以通过对所有候选人的回答进行评分并挑选最高评分的响应来选择回复。整个过程如图1所示。

当ASR置信度低于一定阈值时，系统请求用户重复上一次发言。否则，系统不会使用ASR信任。特别地，响应模型和*模型选择策略*都不使用ASR信任。然而，ASR系统远非完美。因此，通过对ASR信任的响应模型和*模型选择政策*进行调整，可以显着改善系统的可行性。

3        响应模型

系统中有22个响应模型，包括基于检索的神经网络，基于生成的神经网络，知识库问答系统和基于模板的系统。候选模型响应的示例如表1所示。本节旨在总结这些响应模型。首先，描述基于模板的模型。

     3.1      基于模板的模型

我们首先在系统中描述基于模板的响应模型。

Alicebot：*Alicebot*使用一组AIML（人造智能标记语言）模板，以产生对话历史和用户话语的回应（Wallace 2009，Shawar＆Atwell 2007）。我们使用免费提供的Alice内核，[网址](http://www.alicebot.org/)为[www.alicebot.org 。](http://www.alicebot.org/)默认情况下，所有模板都会生成非优先级响应，因此我们可以配置与社交名称，年龄和位置相关的模板以输出优先级响应。我们进一步修改几个模板，使其与挑战一致（例如，避免淫秽语言，并鼓励用户讨论某些主题，如新闻，政治和电影）。大多数模板保持不变。

该*Alicebot*模式还输出标信心评分。由于AIML模板重复用户的输入语音，它们并不总是正确的句子。因此，我们使用基于字符串的规则来确定响应是否构成正确的句子。如果响应是正确的句子，它返回一个很高的表1：响应模型生成的示例对话和相应的候选响应。最终系统的响应标记为粗体。

|  |  |  |  |
| --- | --- | --- | --- |
| 对话 | | 对话 | |
| 系统：嗨 这是Alexa Prize社交人。  用户：嗨  系统：你有宠物吗？  用户：是的，我有两只兔子。 | | ...  用户：我们来谈电影。  系统：你最喜欢的电影是什么？用户：星球大战电影 | |
| 候选人回应 | | 候选人回应 | |
|  | BoWEscapePlan：我没有答案。  BoWFactGenerator：你知道男性的兔子被称为美元，女性是吗？GRUQuestionGenerator：美国家庭怎么样？  发起者：你相信一见钟情吗？  LSTMClassifierMSMarco：2017年3月23日两部分：学习什么饲喂你的兔子学习如何...  DualEncoderRedditNews：bye doggie DualEncoderRedditPolitics：谁是edger你shboys？  爱丽丝：哈里！两只是很多的兔子。  Elizabot：好的，可以吗？VHREDRedditMovies：嗯 不，没听说过。  VHREDSubtitles：你想看看吗？  HelpResponse：我真的想谈谈新闻，政治或电影。 | BoWEscapePlan：我没有答案。BoWFactGenerator：这是一个有趣的事实。南极国际电话拨号是672. GRUQuestionGenerator：纽约市怎么样？  发起者：你今天过得如何？  LSTMClassifierMSMarco：第三部Anthology影片将在2020年发行。  DualEncoderRedditNews：现在耳语对我来说，还有一次是bby。  DualEncoderRedditPolitics：哈哈哈在移动这个甚至没有接近。我希望我可以告诉你，但这看起来很糟糕我很抱歉的人。  爱丽丝：你喜欢那部电影中的机器人？  Elizabot：这怎么让你感觉到？  星际大战电影“星球大战”系列中的电影。  VHREDRedditMovies：哦，请。请。Pleeeease。让这种情况发生。  VHREDSubtitles：什么？  HelpResponse：我真的想谈谈新闻，政治或电影。 |  |
|  |  |  |  |

自信，否则它会返回低信任度。该过程在算法1中说明。

算法1：Alicebot

1   输入：对话历史

2   应答← 将AIML模板应用于对话历史记录

3   ，如果*反应是正确的句子*，然后

五

6

7

9

到Elizabot类似*Alicebot*，该*Elizabot*模型进行字符串匹配以选择一组模板的答案。该模型基于着名的Eliza系统，旨在模仿Rogerian心理治疗师。（Weizenbaum 1966）。[[4]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn4" \o ")因此，与*Alicebot*相反，大多数*Elizabot*的回答是个人问题，这意味着让用户继续谈话。

这里有两个示例模板：

1.     *“我是（。\*）”* → *“你来我往，因为你是......”*

2.     *“什么（。\*）”* → *“你为什么问？*

省略号标记响应语句的部分，将用用户的话语替换文本。该模型检测适当的模板并选择相应的响应（如果有多个模板，则随机选择一个模板）。然后，模型通过一组*反射*来运行模板响应，以更好地格式化字符串以进行响应（例如*“我会”* → *“你会”*，*“你的”* → *“我的”*）。

3

五

Initiatorbot *Initiatorbot*模型充当*对话启动器*：它向用户询问一个开放式问题，让对话开始，并增加用户的参与度。我们为*Initiatorbot*写了40个问题短语。短语的例子包括*“你今天做了什么？*，*“你有宠物吗？”*和*“你最感兴趣的是什么样的新闻报道？”*。作为一个特殊情况，该模型也可以通过说明一个有趣的事实来开始对话。在这种情况下，发起者短语是*“你知道那个<fact>吗？*其中*事实*被一个声明所取代。模型，稍后描述。

在返回响应之前，*Initiatorbot*首先检查它在会话的最后两圈中是否还没有被触发。如果用户发出问候（例如*“嗨”*），则*Initiatorbot*将以优先级返回响应。这很重要，因为我们观察到，问候语通常表明会话的开始，用户没有一个他们想要谈论的特定主题。通过提出一个问题，系统*主动*（即控制对话）。该过程在算法2中详细描述。

Storybot *Storybot*模型根据用户的要求输出一个简短的小说故事。我们实施了这个模式，因为我们观察到许多用户要求社会工作者讲故事。[[5]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn5" \o ")*Storybot*确定用户是否请求了一个故事通过检查是否存在既是请求字（例如*说*，*告诉*）和故事型字中的发声（例如*故事*， *故事*）。这个回应讲述了故事的标题和作者，后面是故事的主体。例如，来自这个模型的一组回应遵循模式*“好吧，让我告诉你<story\_author>的故事<story\_title> <story\_body>*其中<story\_title>是故事的标题，<story\_body>是主要文本，<story\_author>是故事作者的名字。故事从网站：[www.english-for-students.com上](http://www.english-for-students.com/)刮下来[。](http://www.english-for-students.com/)

一个例子是：

蚂蚁和蚱蜢\*\*

蚂蚁在夏天努力工作。他们为冬季分拣食物。

那时候，蚱蜢依然闲着。冬天来了，蚂蚁已经吃饱了。

但是，蚱蜢没有什么可吃的。他不得不饿死。

他去了蚂蚁，乞求食物。蚂蚁回答说：“你在夏天做了什么？他回答说：“我在夏天闲置了我的时间”。

蚂蚁回答说：“那你冬天就饿死了。” MORAL：永远不要空闲

该*Storybot*是在系统中执行的唯一组件*非会话活动*。只有当用户专门要求一个故事时才触发，在这种情况下，它的响应是优先级响应。否则，*Storybot*响应模型从不被触发。此外，系统的其余部分不会鼓励用户请求故事。

     3.2       基于知识库的问题回答

Evibot *Evibot*响应模型将用户的话语转发到亚马逊的问答式网络服务*Evi*：[www.evi.com 。](http://www.evi.com/) *Evi*主要用于处理事实问题。因此，*Evibot*为直接问题返回优先级回应，定义为包含wh字词（例如*“who”*，*“what”*）的用户话语，否则返回非优先级或可能为空的响应。如果查询是一个直接的问题，并且包含不间断的单词，*Evibot*将按照三步骤来生成响应。首先，*Evibot*转发查询[www.evi.com](http://www.evi.com/)包含整个用户的话语，如果有效，返回结果。如果失败，*Evibot*将NLTK的命名实体处理器（Bird et *al。2009）*应用于查询以查找具有命名实体的子查询。对于包含命名实体的每个子句，*Evibot*将查询转发到[www.evi.com ，](http://www.evi.com/)并在有效响应时返回结果。最后，如果前两个步骤失败，*Evibot会*转发每个没有命名实体的子查询的查询，并返回一个有效的响应或一个空的响应。该过程在算法3中详细描述。

算法3：Evibot

1      输入：对话历史

2      查询← 最后一个用户话语

3      has-wh-words ← 如果话语包含一个wh字，则为true，否则为false

4      只有停止词← 如果话语只有停止词，否则为false

5      ，如果*有只停字，而不是具有-WH-话*那么

6      回“”

7      evi-response ← 发送查询到[www.evi.com](http://www.evi.com/)

8      优先← 如果has-wh-words和evi-response有效，则为true，否则为false

10

12

13

14

15

16

17

18

19

20 21

22

23

BoWMovies *BoWMovies*模型是一个基于模板的响应模型，它处理电影领域的问题。该模型具有实体名称和标签的列表（例如*电影情节*和*发行年份）*）。该模型搜索用户的话语中已知的实体和标签。实体通过字符串匹配来标识。这是以级联顺序完成的，通过给电影标题匹配，然后演员名称匹配，最后导演名称匹配的首选。标签也通过字符串匹配来标识。然而，如果标签的精确字符串匹配失败，则通过字嵌入相似性进行标识。如果存在实体和标签，则代理将向多个数据源之一发送API调用，以检索所选查询类型的数据项。代理商受到其可访问的API中可用的数据的限制。模型的响应遵循预定义的模板。

电影名称，演员姓名和导演姓名从互联网电影数据库（IMDB）中提取。

电影说明来自Google Knowledge Graph的API。其他电影标题查询指向开放电影数据库（OMDB）。[[6]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn6" \o ")对于演员和导演查询，使用了Wikiedata API。首先，在Wikidata JSON转储中搜索actor和director名称。

如前所述，该模型使用单词嵌入来匹配标签。这些词嵌入是使用Word2Vec在IMDB数据库提取的电影情节摘要和演员传记中训练的（Mikolov等人，2013）。

算法4：BoWMovies - ComputeResponse

1   输入：对话历史

2   实体← 实体所包含无论是在过去的使用者说话和电影片名，演员或导演的名单3 ，如果*没有实体，*然后

6

8

10

12

算法5：BoWMovies - ComputeEntityResponse

1   输入：实体和实体类型

2   标签← 字符串匹配标签，其中标签对实体类型有效（电影标题，演员名称，导演名称）

3   如果*没有标签，*那么

4 标签← 字嵌入匹配标签，其中标签是单字，对于实体类型有效（电影名称，演员姓名，导演姓名）

5 如果*没有标签，*那么

6 标签← 字嵌入匹配标签，其中标签是多个字，对实体类型有效（电影名称，演员姓名，导演姓名）

7 如果*没有标签，*那么

8 回“”

9      api-response ← 调用外部API与查询（实体，标签）。

10    反应← 插入api响应的模板

11    回应

     3.3       基于检索的神经网络

VHRED模型：该系统包含几个VHRED模型，具有高斯潜在变量的序列到序列模型被训练为变分自动编码器（Serban等人，2017，Kingma＆Welling 2014，Rezende等人，2014）。使用与Serban等人相同的程序训练模型。（2017年）。训练有素的VHRED模型产生候选响应如下。首先，使用词汇TF-IDF手套词嵌入（Pennington et al。，2014），从数据集中检索出一组*K*模型响应，该数据集使用当前对话历史和数据集中的对话历史之间的余弦相似度进行检索。[[7]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn7" \o ")通过VHRED计算每个20个响应的对数似然度的近似值，并返回具有最高对数似然度的响应。该系统具有4种基于Reddit的数据集VHRED模型，一种基于新闻文章的VHRED模型和一种基于电影字幕的VHRED模型：

•     *VHREDRedditPolitics*对[https://www.reddit.com/r/politics进行了](https://www.reddit.com/r/politics)培训，并从*K* = 10的所有Reddit数据集中提取响应，

•     *VHREDRedditNews*培训了Reddit [https://www.reddit.com/r/news，](https://www.reddit.com/r/news)并从*K* = 20的所有Reddit数据集中提取响应，

•     *VHREDRedditSports*在Reddit <https://www.reddit.com/r/sports>上*进行*培训，并从*K* = 20的所有Reddit数据集中提取响应，

•     *VHREDRedditMovies*在Reddit <https://www.reddit.com/r/movies>上训练并从*K* = 20的所有Reddit数据集中提取响应，

•     *VHREDWashingtonPost*[[8]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn8" \o ")在Reddit <https://www.reddit.com/r/politics>上进行了培训，并从用户评论中回复了华盛顿邮报的新闻文章，以及

•     *VHREDSubtitles*[[9]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn9" \o ")使用电影字幕数据集SubTle（Ameixa et al。2014）with

*K* = 10 。

特别地，*VHREDRedditPolitics*和*VHREDWashingtonPost*使用不同的检索过程。这两个模型使用逻辑回归模型来评估响应而不是近似对数似然。对Logistic回归模型进行了一系列的7500条线索训练，亚马逊机械工匠在Likert型规模1-5。根据余弦相似度，手势词嵌入，候选响应从其他Reddit线程中选择。逻辑回归模型的标签收集和训练程序与第4节中描述的程序相似。对于每个响应，逻辑回归模型将VHRED对数似然分数以及其他几个输入特征作为输入，输出一个标量值得分。尽管逻辑回归模型确实提高了为Reddit线程选择的响应的适当性，但*VHREDRedditPolitics*在最终系统中极少使用（参见第4节）。这表明，根据标记的Reddit线程和响应，培训一个模型以重新排列响应不能帮助提高性能。

SkipThought矢量模型：该系统包含在BookCorpus数据集（Zhu et al。2015）和SemEval 2014 Task 1（Marelli et al。，2014）上训练的SkipThought Vector模型（Kiros et al。，2015）。使用与Kiros等人相同的程序训练该模型。（2015），被称为*SkipThoughtBooks*。

*SkipThoughtBooks*确保系统符合亚马逊Alexa奖比赛规则。*竞争早期介绍的一个规则是，社会团体不应该陈述自己与政治或宗教话题有关的意见。如果用户希望讨论这些话题，社会团体应该通过提出问题或说明事实来进行。SkipThoughtBooks*还处理Alexa平台特有的特殊问题。*例如，许多用户不明白社会机构的目的，并要求我们的社交媒体播放音乐。*在这种情况下，系统应该指示用户退出*社交应用*，然后播放音乐。

*SkipThoughtBooks*遵循两步过程来生成其响应。*第一步将用户的最后一个话语与一组触发短语进行比较。如果找到匹配项，模型返回相应的优先级响应。*[[10]](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftn10)例如，如果用户说*“你怎么看待唐纳德王牌？”*，模型将返回一个优先响应，如*“有时，真相比小说更陌生”。*。*找到匹配，如果：1）SkipThought向量模型在用户的最后一个发音和触发短语之间的语义相关性得分高于预定阈值，以及2）用户的最后一个话语包含与触发短语相关的关键字。*[[11]](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftn11)总体上，有315触发短语（大多数是彼此的释义）和35个响应集。

如果模型在第一步没有找到匹配项，则继续进行第二步。在此步骤中，该模型从所有Reddit数据集响应中选择其响应。如前所述，使用余弦相似性来检索一组*K个*模型响应。然后，模型返回具有最高语义相关性得分的响应。

双编码器型号：该系统包含两个双编码器检索模型（Lowe et *al。2015*），*DualEncoderRedditPolitics*和*DualEncoderRedditNews*。两个模型由两个序列编码器ENC *Q*和ENC *R组成*，具有单个LSTM复用层，用于编码对话历史和候选响应。通过对话历史嵌入和候选响应嵌入的双线性映射计算候选响应的分数，如Lowe等人。（2015年）。

使用（Lowe等人，2015）提出的方法训练模型。使用基于手势词嵌入的TF-IDF余弦相似度来检索来自一组*K* = 50 候选响应的最高分数的响应。*DualEncoderRedditPolitics*模型在Reddit <https://www.reddit.com/r/politics> 数据集上进行了培训，并提取了所有Reddit数据集的响应。*DualEncoderRedditNews*模型在Reddit <https://www.reddit.com/r/news>数据集上进行了培训，并提取了所有Reddit数据集的响应。

Bag-of-Words检索模型：该系统包含基于TF-IDF手套字嵌入（Pennington等人，2014）和Word2Vec嵌入（Mikolov等人，2013）的三个袋子检索模型。[[12]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn12" \o ")类似于VHRED模型，这些模型检索具有最高余弦相似度的响应。该*BoWWashingtonPost*模型使用检索词手套的嵌入来自华盛顿邮报的新闻文章的评论。*BoWTrump*模型从Donald Trump的个人资料：[https](https://twitter.com/realDonaldTrump) : *//twitter.com/realDonaldTrump中查到*一组Twitter推文的响应[。](https://twitter.com/realDonaldTrump)该模型还使用手套字嵌入，并且只有在用户的话语中找到至少一个相关关键字或短语时，才会返回响应（例如，当单词用户提到*“Trump”*）。触发关键字和短语的列表包括：*“唐纳德”*，*“王牌”*，*“POTUS”*，*“美国总统”*，*“美国总统”*，*“希拉里”*， *“克林顿”*，*“奥巴马”*，和*'奥巴马“*。模型*BoWFactGenerator*从一组约2500个*有趣*和*有趣的*事实中获取响应，包括关于动物，地理和历史的事实。*该模型使用Word2Vec字嵌入。BoWGameofThrones*模型检索从一组报价中提取的响应[https://twitter.com/ ThroneQuotes](https://twitter.com/ThroneQuotes)使用手套字嵌入。[手动检查和清理来自此源的推文，以删除该系列中没有引号的任何推文。](https://twitter.com/ThroneQuotes)与*BoWTrump*模型一样，我们使用一个触发短语列表来确定模型的输出是否与用户的话语相关。[我们填写了这个列表，其中包含大约80个流行的字符名称，地名和姓氏，这些域名是唯一的。我们还添加了一些别名来尝试和解释这些命名实体的替代语音转录。](https://twitter.com/ThroneQuotes)一些短语包括：*“ned stark”*，*“jon snow”*，*“john snow”*，*“samwell tarly”*， *“hodor”*，*“dothraki”*等等。[[13]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn13" \o ")

     3.4        基于检索的逻辑回归

BoWEscapePlan：系统包含一个响应模型，称为*BoWEscapePlan*，它返回一组35个与主题无关的通用预定义响应的响应，例如*“可以重复一遍”*，*“我不知道”*和*“那是个问题吗？“*。其主要目的是保持用户参与，并保持对话，其他模式无法提供有意义的回应。该模型使用逻辑回归分类器基于一组较高级别的特征来选择其响应。

训练逻辑回归分类器，我们注释12 *，* 000 为妥当用户话语和候选响应对在李克特式量表1 - 5 。用户的话语是从Alexa用户和系统的初步版本之间的交互中提取出来的。从*BoWEscapePlan*随机抽样候选人的响应列表。逻辑回归模型的标签收集和训练程序与第4节中描述的程序相似。逻辑回归模型训练训练集上的对数似然，并在开发集上早期停止，并对测试进行评估组。然而，训练有素的模型的表现很差。得到Pearson相关系数为0 *。*05 和Spearman的秩相关系数为0 *。*07。这表明，与选择均匀随机的响应相比，逻辑回归模型在选择与主题无关的通用响应方面略好一点。未来的工作应该调查收集更多的标签数据和预训练逻辑回归模型。

     3.5        基于搜索引擎的神经网络

该系统包含一个深层分类器模型，称为*LSTMClassifierMSMarco*，它从一组搜索引擎结果中选择其响应。系统使用最后一个用户话语作为查询搜索Web，并检索前10个搜索片段。检索的片段通过剥离尾随字来预处理，删除不必要的标点符号并截断到最后一个完整句子。该模型使用双向LSTM将最后的对话话语和片段分别映射到自己的嵌入向量。将得到的两个表示被连结并通过一个MLP来预测之间的标量值0 - 1 指示片断如何适当是作为对发声的响应。

该模型作为二进制分类模型被训练在具有横截面的Microsoft Marco数据集上以预测给定用户查询的片段的相关性（Nguyen等人，2016）。给定一个搜索查询和一个搜索片段，当搜索片段相关且否则为零时，该模型必须输出一个。搜索查询和基本真实搜索片段被视为正样本，而其他搜索片段随机选择为阴性样本。在这个任务上，模型能够达到72 的预测精度*。*96％的微软Marco开发集。

该系统能够使用来自各种搜索引擎的搜索API，包括Google，Bing和

AIFounded（Im 2017）。在目前的模式中，我们选择谷歌作为搜索引擎，因为定性检查显示，这是最适合的答案。

     3.6       基于生成的神经网络

该系统包含一个生成型*循证*神经网络语言模型，称为*GRUQuestionGenerator*，可以*逐个*产生后续问题，以对话历史为条件。模型的输入由三个组成部分组成：当前单词的一个热向量，一个二进制问题标签和一个二进制的扬声器标签。该模型包含两个GRU层（Cho et al。2014）和softmax输出层。该模型是在Reddit政治和Reddit新闻对话中进行培训，其中通过检测问号将帖子标记为问题。我们使用优化器Adam（Kingma＆Ba 2015），并通过检查验证集的困惑来执行早期停止。对于代，我们首先在一个简短的问题模板（例如*“如何”）中*对模型进行处理，

*“怎么样”*，*“你怎么想”*，*“你的意见是什么”*），然后通过从模型中抽取出问题的标签，将剩下的问题产生出来。*一旦检测到问号，生成过程就停止。此外，通过调整softmax层的温度来控制问题的长度。由于速度要求，仅生成两个候选响应，并返回前10个字的最佳一个对数似然。*

4        型号选择政策

在生成候选响应集之后，对话系统使用*模型选择策略*选择它返回给用户的响应。系统必须选择一个响应，增加用户对整个对话的满意度。它必须在即时和长期的用户满意度之间取得平衡。例如，假设用户要求谈论政治。如果系统选择以政治笑话回应，用户可能会乐意转一圈。然后，用户可能会对系统无法辩论政治问题感到失望。相反，如果系统选择回应一个简短的新闻故事，用户可能不太喜欢一轮。但是，新闻故事可能会影响用户跟进事实的问题，系统可能会更好地处理。为了在即时和长期用户满意度之间进行权衡，我们考虑选择适当的响应*顺序决策问题*。*本节介绍了学习模型选择政策的五种方法。这些方法都将在下一节中与真实用户进行评估。*

我们使用强化学习框架（Sutton＆Barto 1998）。该系统是一个代理，它在环境中采取行动以最大限度地提高奖励。对于每个时间步长*t* = 1 *，...，T*，代理观察对话历史*h t，*并且必须选择*K个*动作（响应）之一：。采取动作之后，代理接收奖励*ř 吨*并转移到下一个状态*ħ 吨*1（其包括在用户的下一个响应）。然后，向代理提供一组新的*K*动作：

。代理商的目标是最大限度地提高奖励的折扣：

*Ť*

*[R* = X*γ 吨 ř 吨，*（1）

*t* = 1

这被称为*预期累积回报*（或简单地预期回报）。参数*γ*＆Element;（0 *，* 1] 是折扣因子。

我们设置的一个问题是，一组动作根据状态（对话历史）而变化。这是因为候选人的回应是由响应模型产生的，这也取决于对话历史。另外，响应模型不是确定性的。这意味着每次代理遇到相同的状态*h t*时，候选响应的集合可能是不同的。[[14]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn14" \o ")这与某些强化学习问题形成了对比，例如学习玩阿塔里2600场比赛，其中一组动作是根据状态而固定的。为了简化符号，我们将把今后的行动数量修正为*K。*

行动价值参数化：我们使用两种不同的方法来参数化代理人的政策。第一种方法是基于由参数*θ*定义的动作值函数：

                                                                      对于 *k*= 1 *，...，K，*（2）

其中估计采取行动的预期收益（候选响应*ķ*给对话历史）*^ h 牛逼*并给予代理将继续使用后相同的策略。鉴于*Q θ*，代理人选择具有最高预期收益的作用：

*π θ*（*ħ 吨*）= argmax *Q θ*（*ħ 吨，一个ķ吨*）*。*（3）

*ķ*

使用动作值函数来选择对话响应与Lowe等人最近的工作密切相关。（2017），其中有一个模型被学习来预测对话系统响应的质量。然而，在我们的情况下，*Q θ*仅空调的对话环境。另一方面，Lowe等人提出的模型 （2017年）受到对话背景和人类参考响应的制约。动作值函数也与Yu等人的工作有关。（2016），他们学习评估模型，用于训练强化学习代理人来选择适当的对话应对策略。

随机策略参数化：第二种方法将策略参数化为对动作的离散分布。设*θ*为参数。代理人通过抽样选择其行动：

                                                        对于 *k*= 1 *，...，K，*（4）

其中*˚F θ*（*ħ 吨，一个ķ吨*）是*评分函数*，其标量分数分配给每个响应*一ķ吨*给定*ħ 吨*。参数*λ*称为温度，并控制分布的熵。*λ*越高，选择动作越均匀。通过选择最高概率的行动，可以将随机政策转化为确定性（贪心）政策：

                                           ）= argmax       ）= argmax *。*（5）

*k                                                            k*

评分模型：动作值函数 和评分函数 密切相关。这两个功能都是对行动的排名; 较高的价值意味着更高的预期回报。

当，等式中的动作值函数策略 （3）相当于方程式中的贪心政策。（5）。为简单起见，我们将使用两个相同的参数化 和*˚F θ*（*ħ 吨，一个ķ吨*）。因此，我们让两个函数具有与输入相同的特征，并使用相同的神经网络架构处理它们。我们将这两个功能称为*评分模型*。

下一节将介绍评分模型的输入特征。

     4.1       输入功能

作为评分模型的输入，我们基于给定的对话历史和候选响应计算1458个特征。输入特征是基于词嵌入，对话行为，词性标签，单字重叠，双字重叠和模特特征的组合。

        响应字的嵌入：                       平均    的    候选    响应   字     embed-

（Mikolov等人，2013）。[[15]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn15" \o ")

最后一个用户话语的词嵌入：最后一个用户发音字嵌入的平均值。

上下文                             的词嵌入：对话语境中最后六个话语的嵌入词的平均值。

|  |  |
| --- | --- |
| 用户上下文的字嵌入 | 在对话上下文中最后三个用户话语的嵌入词的平均值。 |
| 词嵌入相似性度量： | 该*嵌入平均*，*嵌入极值*和  Liu等人描述的*嵌入贪婪*相似性度量 （2016）。*每个相似性度量在1）最后一个用户话语和候选响应之间计算，2）对话和候选响应中的最后六个话语，3）对话和候选响应中的最后三个用户话语，4）最后六个话语对话和候选人的回应与停止词被删除，以及5）最后三个用户话语在对话和候选人回应与停止词被删除。* |
| 响应模型类： | 大小等于响应模型数量的一个热向量，其中条目*i*等于1 *。*当候选人的回应由具有索引*i*的模型类生成时，为0 。 |
| 词性反应类： | 候选响应的词性标签是使用Penn Treebank语料库训练的最大熵标签来估计的。然后将部分语音标签的序列映射到构成输入特征的单热向量。 |
| 对话行为反应模式类： | 表示对话行为的一个热向量（我们考虑10种对话行为）和用于指示模型类的一个热向量之间的外部乘积（Stolcke等人，2000）。 |
| 词重叠： | 1 *。*0，当一个或多个不间断词在候选响应和最后一个用户话语之间重叠时，否则为零。 |
| Bigram重叠短期： | 1 *。*在候选响应和最后一个用户话语中都存在一个二进制（两个连续的令牌）时为0 ，否则为零。 |
| Bigram长期重叠： | 1 *。*0 时两字都在候选响应，并在对话上下文中的最后话语之一，否则为零存在。 |
| 命名实体重叠短期： | 1 *。*0，当候选响应和最后一个用户话语中存在命名实体（不是停止字的上位字），否则为零。 |
| 命名实体重叠长期： | 1 *。*0 时，一个名为实体都在候选响应，并在对话上下文中的最后话语之一，否则为零存在。 |
| 一般回应： | 1 *。*0，当候选人的回应只包括不到3个字符的词或短语，否则为零。 |
| Wh字词回应功能： | 1 *。*0 时候选响应中包含一个疑问词（例如  *什么*，*在哪里*等等），否则为零。 |
| 字词上下文： | 1 *。*0，当最后一个用户发音包含一个wh字，否则为零。 |
| 增强词反应： | 1 *。*0，当候选响应包含增强词（例如*令人惊讶*，*疯狂*等等），否则为零。 |
| 增强词上下文： | 1 *。*0，当最后一个用户发音包含增强字，否则为零。 |
| unigram回应： | 一组二进制特征是1 *。*0 当候选人回应包含一个特定的单词（包括单词*I*，*你*和*谢谢*），否则为零。 |
| 否定回应： | 1 *。*0 时候选响应中包含一个否定词，如*不*或*不*，否则为零。 |
| 不间断的回应： | 1 *。*0 当候选响应包含非停止词，否则为零。 |

由于实验原因，我们不包括基于语音识别系统的信心的功能。语音识别错误是真实用户实验中的混淆因素。语音识别错误可能会影响用户满意度。如果包含基于语音识别信息的功能，则一种策略可能会比另一种策略更好地处理语音识别错误。反过来，这可能会使该策略的整体用户满意度更好。然而，这将是由不完美的语音识别系统引起的影响，并不会在完美的语音识别系统下反映用户满意度。排除这些特征作为评分模型的输入有助于最小化这种混杂效应。然而，即使排除了这些特征，应该注意的是，语音识别误差在后面的实验中仍然构成了实质的混杂因素。最后，出于同样的原因，没有一个响应模式利用语音识别信息。

原则上，可以通过使用递归神经网络（RNN）或卷积神经网络（ConvNets）编码对话上下文和候选响应来计算输入特征（Socher等人，2013，Blunsom等人，2014，Cho等人2014 ，Yu et al。，2014，Kiros et al。，2015）。然而，这些模型被认为需要对大型语料库进行培训，才能实现可接受的性能，我们无法访问。此外，我们需要将得分模型的执行时间保持在150ms以下。否则，响应时间的放缓可能会使用户失望，降低整体用户满意度。这排除了亚马逊Alexa奖大赛的大型RNN和ConvNets，因为这将需要更多的计算运行时间。然而，使用较大数据集的未来对话系统应考虑大规模模型。

     4.2       模型架构

本节介绍评分模型的架构。评分模型是一个五层神经网络。第一层是输入，由上一节描述的1458个功能组成。第二层包含500个隐藏单位，通过将线性变换和纠正的线性激活函数（Nair＆Hinton，2010，Glorot等人2011）应用于输入层单元来计算。第三层包含20个隐藏单位，通过对前一层单位应用线性变换计算。类似于矩阵分解，该层将500个隐藏单元压缩到20个隐藏单元。第四层包含5个输出单位，它们是概率（即所有值都为正，并且总和为1）。这些输出单元是通过对前面的层单元应用线性变换，然后进行softmax变换来计算的。该层对应于Amazon Mechanical Turk标签，这将在下一小节中进行描述。第五层是通过对第三层和第四层中的单元应用线性变换来计算的最终输出标量。该模型如图2所示。

在结算这个架构之前，我们尝试了更深层次和更浅的模型。然而，我们发现更深的模型和更浅的模型表现更糟。然而，未来的工作应该探索替代架构。

我们使用五种不同的机器学习方法来学习评分模型。这些将在下面描述。

     4.3          监督AMT：使用群众标签学习

本节介绍了第一种学习扫描模式的方法，该模式基于来自众包示例的监督学习。这种方法也用作稍后讨论的方法的初始化。

众包：我们使用Amazon Mechanical Turk（AMT）收集数据来训练评分模式。我们遵循类似于Liu等人的设置 （2016）。我们向人类评估人员展示一个对话以及4个候选人的回答，并要求他们评估每个候选人的回应是否合适

图2：评分模型的计算图，用于基于动作值函数和随机策略参数的模型选择策略。该模型包括具有1458个特征的输入层，具有500个隐藏单元的隐藏层，具有20个隐藏单元的隐藏层，具有5个输出概率的softmax层（对应于第4.3节中的五个AMT标签），以及标量 -有价值的输出层。虚线箭头表示跳过连接。

1-5 Likert型秤。得分1表示响应不合适或没有意义，3表示响应可以接受，5表示响应非常好，非常适合。

我们的设置只要求人类评估人员评估候选人的反应总体适当性。原则上，我们可以选择评估候选人回应的其他方面。例如，我们可以评估流畅度。然而，流利程度评估不会非常有用，因为我们的大多数模型都从现有的语料库中回答了他们的回答，主要包含流利和语法正确的答案。另一个例子，我们可以评估主题相关性。然而，我们选择不评估这样的标准，因为已知难以达成高标注协议（Liu et al。2016）。事实上，众所周知，即使要求单一的总体评级往往只能在人类评估者之间产生公平的协议（Charras et al。2016）;

对话是从Alexa用户和我们系统的初步版本之间的交互中提取出来的。仅提取系统没有优先级响应的对话（当有优先级响应时，对话管理器必须始终返回优先级响应）。约3 */* 4 这些对话后，随机抽出，剩下的1 */* 4 的对话，随机不含相同对话采样。[[16]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn16" \o ")对于每一对话，相应的候选响应由从所述响应模型候选响应创建。

我们通过掩饰亵渎和用星星发誓的话来预处理对话和候选人的回答（例如我们将*“他妈的”*映射到*“\*\*\*\*”*）。[[17]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn17" \o ")此外，我们通过以随机选择的性别中立名称替换名字来匿名对话和候选答复（例如，*“Hi John”*可以映射到*“Hello Casey”*）。最后，对话被截断到最后4个话语和最后500个单词。这降低了注释器的认知负荷。众包任务的示例如图3，图4和图5所示。图5所示的对话示例是一个虚构的例子。

图3：Amazon Mechanical Turk人类智能任务（HIT）的同意屏幕。

图4：Amazon Mechanical Turk人类智能任务（HIT）的说明屏幕。

图5：Amazon Mechanical Turk人类智能任务（HIT）的注释屏幕。对话文本是一个虚构的例子。

我们手动检查了注释。我们观察到，注释者倾向于频繁地超过与主题无关的通用响应。这样的回答可能被认为是可以接受的一个单一的谈话中的转弯，但可能是有害的，当重复一遍又一遍。特别是，注释者倾向于超过响应模型*Alicebot*，*Elizabot*，*VHREDSubtitles*和*BoWEscapePlan*生成的响应。这些模型生成的响应通常是可接受的或很好的，但是它们中的大多数是与主题无关的通用句子。因此，对于这些响应模型，我们将所有标签5 （*“优”*）映射到4 （*“好”*）。此外，对于仅包含停止词的响应，我们将标签减少一个级别（例如，4 被映射到3 ）。最后，*BoWMovies*响应模型在标签收集期间遭遇了一个错误。因此，我们将所有给予*BoWMovies*响应的标签减至*最少* 2 （*“差”*）。

我们总共收集到199 *，* 678 的标签。我们分为培训（火车），开发（dev）和测试（测试）数据集，分别包括137,549,23,298和38,831个标签。

培训：我们优化评分模型对数似然（交叉熵）来预测代表AMT标签类的第4层。正式地，我们优化参数*θ*：

*θ*= argmax X日志*P θ*（*Ý* | *X*）*，*（6）*θ的x，y*

其中*X*是所述输入功能，*ÿ*是对应的AMT标签类（一热载体）和*P θ*（*Ý* | *X*）是该模型的预测的概率*ÿ*给定*X*，在评分模型的第二最后一层计算。我们使用一阶梯度下降优化器Adam（Kingma＆Ba 2015）我们尝试了各种超参数，并根据开发组的对数似然度选择最佳的超参数组合。对于第一隐藏层，我们用层尺寸实验在集合： {500 *，* 200 *，* 50}。对于第二个隐藏层，我们在与集层的大小进行实验：{50 *，* 20 *，* 5} 。我们对所有模型参数使用L2正则化，偏置参数除外。我们对集合中的L2正则化系数进行实验：{10 *。*0 *，* 1 *。*0 *，* 10 -1*，...，* 10 -9 } 不幸的是，我们没有标签训练的最后一层。因此，我们将最后一层的参数修改为向量

[1 *。*0 *，* 2 *。*0 *，* 3 *。*0 *，* 4 *。*0 *，* 5 *。*0] 。换句话说，我们分配一个得分的1 *。*0 为标签 *非常差*，一分

2 *。*0 为标签 *差*，得分为3 *。*0 标签*可接受*，得分为4 *。*0 标签 *好*，得分5分*。*0 为标签*优秀*。由于此模型是针对Amazon Mechanical Turk（AMT）的众包数据进行了培训的，所以我们将此模型称为*监督AMT*。

表2：亚马逊Mechanical Turk测试集的评分模型评估Pearson相关系数，Spearman秩相关系数和均方误差。

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 皮尔逊 | 斯皮尔曼 | 均方误差 |
| *平均预测因子* | 0 *。*00 | 0 *。*00 | 1 *。*三十 |
| *监督AMT* | **0***。***40** | **0***。***38** | **1***。***10** |

图6：Amazon Mechanical Turk类测试集上的频率不同策略。

表2显示了Pearson相关系数，Spearman秩相关系数和均方误差的性能。的度量之后线性变换的AMT类分类标量输出得分（通过取一个热类矢量和矢量之间的点积，即计算的[1 *。* 0 *，* 2 *。* 0 *，* 3 *。* 0 *，* 4 *。* 0 *，* 5 *。* 0] ）。的*平均预测*是基准模型，它总是与输出平均评分预测。如图所示，*监督AMT*实现了Pearson相关系数为0*。*40 ，Spearman的秩相关系数为0 *。*38 和均方误差的显着降低。这表明受*监督的AMT*表现明显优于基线。

图6显示了每个AMT标签类的性能。除了受*监督的AMT*，该图还显示了三个基准策略的性能：1）*随机*，*随机*选择一个响应，2）*Alicebot*，它可以选择一个 *Alicebot*响应，否则随机选择一个响应，3）*Evibot + Alicebot*，它可以选择*Evibot*响应（如果可用），否则选择*Alicebot*响应。对于每个策略，该图显示属于特定AMT标签类的策略选择的响应的百分比。在一方面，我们观察到*监督AMT*有*〜30*％相对于属于*“非常差”*类的响应中的*随机*点减少。对于相同的AMT标签类，与*Alicebot*和 *Evibot + Alicebot*相比，*受监督的AMT*减少了约10％。在另一方面，我们观察到

*受监督的AMT*表现明显优于*“良好”*和*“优秀”*课程的三个基线。特别是*AMTI*达到〜8 ％，属于*“优秀”*类。*与所有三个基准政策相比，这是两倍多。*这表明*监督*

*AMT*已经学会选择*“好”*和*“优秀”的*回应，同时避免*“非常贫穷”*和*“差”的*回应。

总体而言，结果显示，*监督AMT*比所有基准政策大幅改善。然而，46％的*监督AMT*回应属于*“非常贫穷”*和*“穷人”*阶层。这意味着有足够的空间来改善*监督AMT*和候选人回应（即系统的响应模型）。

     4.4受          监督的学习奖励：学习奖励学习功能

在第一个评分模型*监督AMT中，*我们将最后一个输出层权重修改为[1 *。*0 *，* 2 *。*0 *，* 3 *。*0 *，* 4 *。*0 *，* 5 *。*0] 。换句话说，我们分配了1分*。*0 反应*非常差*，2 *。*0 反应*差*，3 *。*0 为*可接受*回应等等。目前还不清楚这个分数是否与真实世界的Alexa用户给出的分数相关，这是我们最终希望优化系统的。本节介绍另一种方法，通过学习基于先前记录的对话来预测Alexa用户分数来补救此问题。

据悉奖励功能：让*^ h 牛逼*是一个对话的历史，并让*一个牛逼*是相应的响应，在时间由系统给出的*牛逼*。我们的目标是学习线性回归模型，*摹φ*，它预测在当前对话又将相应的回报（Alexa的用户评分）：

*克φ*（*ħ 吨，一吨*）∈[1 *，* 5] *，*（7）

其中*φ*是模型参数。我们称之为*奖励模式*，因为它直接模拟了Alexa用户分数，我们旨在最大化。

令{ *h dt，a dt，R d* } *d，t*为一组示例，其中*t*表示时间步长，*d*表示对话。让*ř ð* ∈[1 *，* 5] 表示用于对话所观察到的实值返回*ð*。

具体来说，我们将*R d*设置为对话结束时给出的Alexa用户分数*d*。用户可以选择一个分数; 提示用户在最后给出分数，但是可以通过停止应用程序来选择退出。虽然不是所有的用户都给出分数，但我们不考虑没有分数的例子。[[18]](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftn18)此外，鼓励用户提供1 - 5 的分数。广大用户的给整数（整数）的分数，但一些用户提供十进制分数（例如3 *。* 5 ）。因此，我们将*R d*视为1 - 5 范围内的实数。

我们通过最小化模型预测与观察到的回报之间的平方误差来学习*φ*：

φ = argmax XX（*克φ*（*ħ ð吨，一个ð吨*） - *[R ð*）2（8）

*φ                  ð          吨*

与之前一样，我们使用Adam对小批次随机梯度下降（SGD）优化模型参数。我们使用集合{10 中的系数的L2正则化*。*0 *，* 1 *。*0 *，* 0 *。*1 *，* 0 *。*01 *，* 0 *。*001 *，* 0 *。*0001 *，* 0 *。*00001 *，* 0 *。*0} 。我们在保留数据集上选择具有最小平方误差的系数。

作为奖励模型的输入，我们基于对话历史和候选响应计算23个特征。由于培训数据很少，我们只使用更高层次的功能：

|  |  |
| --- | --- |
| AMT标签类： | 指示使用*监督AMT*计算的候选响应的AMT标签类别的概率的矢量以及候选响应优先的概率。如果候选响应具有优先级，除了对应于优先级的最后一个条目，所有条目中的向量为零：[0 *。*0 *，* 0 *。*0 *，* 0 *。*0 *，* 0 *。*0 *，* 0 *。*0 *，* 1 *。*0] 。 |
| 一般回应： | 二进制特征，即1 *。*0 时响应只包含停止字，否则为零。 |
| 响应长度： | 响应中的单词数，以及响应中单词数的平方根。 |
| 对话行为： | 一个热的向量，指示最后一个用户的话语对话是否是*请求*，*问题*，*声明*或包含亵渎（Stolcke et al。2000）。 |

情绪阶层：                    一个热的向量，指示最后一个用户的话语对话是否为负面，中立或正面。

通用用户话语：二进制特征，即1 *。*0 时，最后一个用户说话只包含停止字，否则为零。

用户话语长度：         最后一个用户话语中的单词数，以及响应中单词数的平方根。

       混乱指标：           二进制特征，即1 *。*0 当最后使用者说话很短

（少于三个字），并且至少包含一个表示用户混淆的单词（例如*“什么”*，*“愚蠢”*， *“愚蠢”*）。

对话长度：到目前为止的对话数量以及对话数的平方根和对数转。

总的来说，我们的培训奖励模型数据集有4340个对话。我们将其分为具有3255个示例的训练集和具有1085个示例的测试集。

为了提高数据效率，我们通过装袋技术的变体来学习一个综合模型（Breiman 1996）。我们创造了5个新的训练集，它们是原始训练集的洗牌版本。每个混洗数据集分为子训练集和子保持集。创建一些集合中的示例不与其他子保持集合重叠。对每个子训练集训练一个奖励模型，其子参数选择其超参数。这可以通过允许我们重新使用辅助组合进行培训来增加数据效率，否则这些组件将不被使用。最终的奖励模式是一个整体，其中输出是底层线性回归模型的平均值。

奖励模型获得均方误差为0 *。*96 和Spearman的秩相关系数为0 *。*19 将真正的Alexa用户放在测试集上。相比之下，用平均用户分数预测的模型获得了0 的均方误差*。*99和（因为它输出一个常数）Spearman的秩相关系数为零。虽然奖励模式优于预测平均值，但其相关性较低。有两个原因。一，培训数据量很小。这使得很难了解功能和Alexa用户分数之间的关系。其次，Alexa用户的分数可能会有很大差异，因为它们受到许多不同因素的影响。用户的得分可以通过对话中的单一回合来确定（例如，即使对话中的所有以前的转换都是出色的），对话结束时的单一误会可能导致非常低的用户评分。用户的分数可能受到语音识别模块的准确性的影响。更多的语音识别错误将不可避免地导致用户沮丧。在初步研究中，我们发现Spearman在语音识别信息与Alexa用户分数之间的秩相关系数在0 *。*05-0 *。*09 。与其他因素的相关性相比，这意味着语音识别性能在确定用户满意度方面起着重要的作用。[[19]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn19" \o ")此外，外在因素可能对用户分数产生重大影响。用户分数可能不仅取决于对话，还可能取决于用户的个人资料（例如，用户是成人还是小孩），环境（例如，用户是否与系统一起使用，或者几个用户正在轮流使用与系统交谈），用户在开始对话之前对系统的期望（例如系统是否能够玩游戏）和用户的情绪状态（例如用户的心情）。

培训：为了防止过度配合，我们不会以奖励模式为目标从头开始训练得分模型。相反，我们首先用*监督AMT*的参数初始化评分模型，然后用奖励模型输出训练，以最小化平方误差：

*θ* = argmaxXX（*˚F θ*（*ħ DT ，一个DT*） - *克φ*（*ħ DT ，一DT*））2 *，*（9）

*θ                ð          吨*

如前所述，我们使用Adam利用随机梯度下降来优化模型参数。由于培训这种模式不依赖于AMT标签，所以在记录对话中进行培训。我们训练了数千个记录的对话例子，其中约80％用于培训，约20％被用作保留套。不使用正则化。我们早点停止了由奖励模式预测的持有数据集wrt Alexa用户分数的平方误差。由于这个计分模式是用学习的奖励功能训练的，所以我们称之为“ *监督学习奖励”*。

     4.5          离职政策强化

如前所述，参数化策略的一种方法是作为离散行为的概率分布。这种参数化使我们能够通过一系列称为*策略梯度*方法的方法直接从记录的对话中学习政策。本节介绍一种这样的方法。

关政策的强化学习：我们使用了经典的变体*加固*算法（威廉姆斯1992年，Precup 2000，Precup等人，2001年），我们称之为*关政策加固*。召回方程 （4），其中在操作的策略的分布被作为参数化函数SOFTMAX施加到函数*˚F θ*与参数*θ*。像以前一样，让{ *h dt，a dt，R d*} *d，t*是一组例子，其中*h dt*是对话的对话历史*d*在代理人的对话行动*d*在时间*t*和*R d*是对话的返回*d*。让*D*是对话的数量，并且令*T d*是对话中的匝数*d*。此外，令*θ ð*是随机策略的参数*π θ 吨*对话期间使用*ð*。该*关政策加固*算法更新策略参数*θ*由：

                Δ *θ* α *Ç ð吨*＆dtri; *θ*登录*π θ*（*一个ð吨* | *ħ ð吨*）*- [R ð*其中*ð*〜统一（1 *，D*）和*吨*〜统一（1 *，T ð*）*，*（10）

其中*c dt*是重要比重：

*。*（11）

该比率校正学习策略之间的差异*π θ*和在其下的数据收集策略*π θ 吨*（有时被称为行为策略）。它在学习政策下具有高概率的权重示例，并且在学习回报函数下具有低概率的下称示例。

算法背后的直觉可以通过与尝试和错误的学习进行类比来说明。当例如具有高的回报（即高的用户得分），术语＆dtri; *θ*登录*π θ*（*一个ð吨* | *ħ ð吨*） *- [R ð*将是一个向量指向的方向上增加采取行动的概率*一个ð吨*。在另一方面，当一个实施例具有低的回波（即，低的用户得分），术语＆dtri; *θ*登录*π θ*（*一个ð吨* | *ħ ð吨*） *ř ð*将是一个向量接近于零或一个矢量指向在相反方向上，因此，降低采取行动的概率*一个ð 吨*。

已知重要性比率*c dt*表现出非常高的，可能无限的方差（Precup等人，2001）。因此，我们将提名者和分母中的产品截断为仅包括当前时间步长*t*：

*c dt，* trunc。 *。*（12）

这在学习过程中引起偏见，但也可以作为正义者。

奖励整形：如前所述，等式1中提出的*非政策性增强*算法存在一个问题。（10）是高度差异（Precup et al。2001）。该算法使用返回值，仅在剧集的最后观察到，以更新策略中的所有中间动作的动作概率。有少量的例子，梯度估计器的差异是压倒性的，这可能容易导致代理人过度估计不良行为的效用，反之亦然，低估了良好行为的效用。这个问题的一个补救办法是*奖励塑造*，其中使用辅助函数估计每个时间步长的奖励（Ng等人，1999）。为了我们的目的，我们提出一个考虑到用户情绪的奖励塑造的简单变体。当用户以负面情绪（例如愤怒的评论）作出回应时，我们将假设前面的操作是非常不合适的，并将其分配给零。鉴于对话*ð*，在每次*牛逼*我们分配奖励*[R 牛逼ð*：



                                                0，        如果用户在时刻*t* + 1的话语有负面情绪，

*ř ð* DEF  *ð*（13）

除此以外。

具有奖励形式和截断重要性权重，学习更新成为：

      其中 *ð*〜制服（1 *，D*）*，T*〜制服（1 *，T ð*）*，*（14）关政策评估：评估政策，我们估计的预期收益（Precup 2000）：

                                                                       [R *。*（15）

*D，T*

此外，用常数奖励1 代替*r td。*0 为每个时间段，我们可以计算估计数的政策下，每集时间步长。如稍后将讨论的，这是一个正交度量，我们可以基于此进行分析和评估每个策略。但是，这个估计不包括优先级响应的数量，因为当存在优先级响应时，代理没有采取任何行动。

培训：我们用*监督AMT*的参数初始化策略模型，然后训练参数wrt eq。（14）随机梯度下降使用亚当。我们使用一组在Alexa用户和系统的初步版本之间记录的几千个对话。其中约60％用于培训，约20％用于开发和测试。为了减少过度配套的风险，我们只使用*非政策性强化*培训与第二层相关的权重。我们使用具有不同超参数的随机网格搜索，其中包括温度参数*λ*和学习率。我们在开发集上选择具有最高预期收益的超参数。

     4.6            具有学习奖励功能的离职政策强化

与“受*监督的学习奖励”*政策类似，我们可以使用*非政策性REINFORCE*算法进行培训的奖励模式。本节介绍如何组合这两种方法。

奖励塑造与学习奖励模型：我们使用奖励模型计算每个对话中每个时间步长奖励的新估计：

*d* def         0                    如果在时间*t* + 1的用户话语具有负面情绪，

*吨*（16）*克φ*（*ħ 吨，一吨*）否则。

这被代入等式 （14）训练和等 （15）进行评估。

培训：与*非政策性REINFORCE一样*，我们用*监督AMT*模型的参数初始化政策模型，然后训练参数wrt eq。（14）使用亚当的迷你批次随机梯度下降。我们使用相同的对话集，并将其拆分为*Off-policy REINFORCE*。我们使用具有不同超参数的随机网格搜索，如前所述，为了降低过拟合的风险，我们仅使用该方法训练与第二层相关的权重。其中包括温度参数*λ*和学习率，并选择具有最高预期回报的超参数。在这种情况下，根据学习的奖励模型计算预期收益。由于这项政策使用了学习的奖励模式，所以我们称之为*非政策性增强学习奖励*。

     4.7         抽象话语马尔可夫决策过程的Q学习

到目前为止描述的方法各有各自的优点和缺点。量化他们差异的一种方法是通过称为*偏差方差折衷*的分解。在频谱的一端，受*监督的AMT*政策方差低，因为它在每个模型响应级别都经过数十万人的注释训练。然而，出于同样的原因，受*监督的AMT*引起了很大的偏见，因为人的注释不能反映真正的用户对整个对话的满意度。另一方面，*离职政策强化*受到高度的差异，因为它只接受了几千个对话和相应的用户分数的训练。更糟糕的是，用户分数受到许多外部因素（例如用户个人资料，用户期望等）的影响，并以整个会话的粒度发生。然而，这种方法引起了低偏差，因为它直接优化了我们关心的客观度量（即用户评分）。[[20]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn20" \o ")通过利用学习的奖励功能，*监督学习奖励*和*非政策性增强学习奖励*受益偏低，但由于学习回报函数有自己的方差分量，所以它们都必然会有较大的差异。一般来说，找出偏差和方差之间的最优权衡是非常困难的。在本节中，我们提出了一种通过在马尔可夫近似马尔可夫决策过程中从模拟中学习政策来消除偏差和差异的新方法。

动机马尔可夫决策过程（MDP）是对顺序决策进行建模的框架（Sutton＆Barto，1998）。在一般设置中，MDP是由离散的状态集合*H*，动作*A*的离散集合，过渡分布函数*P*，奖励分配函数*R*和折现因子*γ组成的模型*。像以前一样，一个代理人的目的是在每一集中最大化其奖励。令*t*表示长度为*T*的情节的时间步长。在时间步长*吨*，所述试剂是在状态*ħ 吨* ∈ *ħ*并采取行动*一吨* ∈*一*。之后，代理接收奖励*ř 吨*〜*- [R*（*ħ 吨，一吨*）并转移到一个新的状态*ħ 吨*1〜*P*（*ħ 吨* | *一吨*）。

考虑到开放域对话的MDP模型，我们可以应用几十种算法来学习代理的策略（Sutton＆Barto 1998）。不幸的是，这样的MDP很难建立或估计。我们可以尝试从记录的对话中天真地估计一个，但这需要解决两个非常困难的问题。首先，我们需要学习过渡分布*P*，在对话历史中，输出下一个用户的话语对话。这个问题很可能像我们原来找到适合用户的一个问题一样困难！其次，我们需要学习奖励分配*R*为每个时间步。然而，正如我们之前所示，很难学习预测整个对话的用户分数。鉴于我们可以获得的数据，估计一轮的回报可能也会很困难。相反，我们建议将问题分解成三个较小的部分。

图7：*抽象话语马尔可夫决策过程的*概率定向图形模型。对于每个时间步长*吨*，*ž 吨*是代表对话的抽象状态的离散随机变量，*ħ 吨*代表对话历史，*一吨*表示由系统所采取的行动（即，选择的响应），*ÿ 吨*表示采样AMT标签和*r t*表示采样的奖励。

抽象话语马尔可夫决策过程我们提出要学习的模型叫做

*抽象话语MDP*。*如图7所示，模型遵循每个时间步长的层次结构。*在时间*吨*，所述试剂是在状态*ž 吨* ∈ *Ž*，代表离散随机变量*抽象话语状态*。*该变量仅代表与对话历史相关的几个高级属性。*我们定义集合*Z*是笛卡尔乘积：

*Z* = *Z*对话行为× *Z*用户情绪× *Z*通用用户话语*，*（17）

其中*ž*对话行为，*Ž*用户情绪和*Ž*通用用户话语是三个离散集。第一集包括10个对话行为：*Z*对话行为 =

{ 接受*，*拒绝*，*请求*，*政治*，*一般问题*，*个人问题*，*声明*，*问候*，*

再见*，*其他} 。这些对话行为代表了用户话语的高级意图（Stolcke et al。2000）。第二组由情绪类型组成：*Z*用户情绪 = { 负*，*中性*，*正面} 。第三组表示一个二进制变量：*Z*Generic user utterance = { True *，* False } 。此变量为*True*只有当用户话语是通用和主题独立的（即当用户话语仅包含停止词）时。我们建立一个手工制作的确定性分类器，将对话历史映射到相应的类别

*Z*对话行为，*Z*用户情绪和*Z*通用用户话语。我们表示这个映射*f h* → *z*。虽然我们只考虑对话行为，情绪和通用话语，但是用其他类型的离散或实值变量来扩展*抽象话语状态*是微不足道的。

给定样本*z t*，*抽象话语MDP*从对话历史*H*的有限集合中对对话历史*h t进行*采样。特别地，*ħ 吨*以随机均匀地从其中最后发声被映射到该组对话历史的采样*ž 吨*：

*ħ 吨*〜*P*（*ħ* | *H，F ħ*→*Ž，Z 吨*）DEF = 统一（{ *ħ* | *ħ* ∈ *ħ*和*˚F ħ*→*Ž*（*ħ*）= *Ž 吨* }） *。*（18）

换句话说，*h t*是对话历史，对话行为，用户情绪和通用属性与离散变量*z t*相同。

为了我们的目的，*H*是Alexa用户和系统的初步版本之间的所有记录对话集。这使得正式的*抽象话语MDP*一个*非参数*模型，因为从模型取样需要访问集录对话的历史*^ h*。该系统在实际部署时随着时间的推移而增长。这是有用的，因为它允许在新数据可用时不断改进策略。此外，应当注意，组*Z*足够小，使得在记录的对话中观察到多次可能的状态。

给定的样品*ħ 吨*，代理人选择的动作*一吨*根据其策略*π θ*（*一吨* | *ħ 吨*），以参数*θ*。奖励*ř 吨*然后进行采样，使得*ř 吨*〜*- [R*（*ħ 吨，一吨*），其中*[R*是一个分布函数。在我们的例子中，我们使用的概率函数*P θ*，其中参数θ估计使用方程式AMT标签上的监督学习。（6）。我们指定了-2 的奖励*。*0 为*“非常差”的*反应班，奖励-1 *。*0 为*“差”*反应班，奖励为0 *。*0 为*“可接受”的*回应班，奖励1 *。*0 为*“好”*反应班和奖励2 *。*0 为*“优秀”*反应班。为了减少超参数，我们使用预期的奖励而不是样本：[[21]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn21" \o ")

*ř 吨* = *P θ*（ *Ý* | *ħ 吨，一吨*）Ť [-2 *。*0 *，* -1 *。*0 *，* 0 *。*0 *，* 1 *。*0 *，* 2 *。*0] *。*（19）

接着，可变*ÿ 吨* ∈{ “非常差” *，* “差” *，* “可接受的” *，* “好” *，* “极好” } 进行取样：

*ý 吨*〜*P θ*（ *Ý* | *ħ 吨，一吨*）*。*（20）

该变量表示输出的一个*适当解释*。这个变量有助于预测未来状态*z t*+1，因为响应的总体适应性对用户的下一个发音有很大的影响（例如，非常差的响应通常会导致用户使用“ *什么”*或“ *我不明白”*进行回应）。

最后，新的状态*ž 吨*1是根据取样*P ψ*：

*ž 吨*1〜*P ψ*（ *ž* | *Ž 吨，H 吨，一吨，Y 吨*）*。*（21）

其中*P ψ*是 带有参数的过渡分布*ψ*。过渡分布由三个独立的两层MLP模型参数化，它们作为输入与评分函数相同的特征，以及1）表示采样响应类别*y t*的单热矢量，2）一热表示最后一个用户话语的对话行为的向量，3）表示最后一个用户话语情绪的一个热向量，4）指示最后一个用户话语是否通用的二进制变量，以及5）二进制变量， 最后一个用户话语包含一个wh字（例如*什么*，*谁*）。第一个MLP预测下一个对话行为，第二个MLP预测下一个情绪类型，第三个MLP预测下一个用户的话语是否是通用的。用于训练的MLP数据集包括499 *，* 757 的转换，它的70％被用于训练和30％进行评估。使用迷你批次随机梯度下降训练MLP的最大对数似然。我们使用亚当和早期停止在一个持有的集合。由于大量示例，不使用正则化。三个MLP模型获得了19 的联合困惑*。*51 。相比之下，总是将平均等级频率分配为输出概率的基线模型获得了困惑23 *。*87 。平均而言，这意味着可以通过对先前变量*z t，h t，a t*和*y t*进行调整来消除大约3 - 4个可能的*z t*+1状态。换句话说，以前的状态*ž ŧ*和*^ h 牛逼*，与代理的一起行动*一个牛逼*对未来状态的显著效果*ž 牛逼*+1。这意味着代理人在*抽象话语MDP中*受过培训在选择其行动时，有可能学会考虑未来的对话状态。这与使用监督学习的政策相反，监督学习不考虑未来的对话状态。

对对话的高级抽象建模*z t*的想法与对话状态跟踪挑战有关（Williams等，2013，2016）。在这个挑战中，任务是将对话历史映射到表示关于对话的所有显着信息的离散状态。然而，与对话状态跟踪挑战不同，变量*z t*仅包括有关对话的有限的显着信息。例如，在我们的实现，*ž 牛逼*不包括主题信息。因此，*z t*只是对话历史的部分表示。

培训鉴于*抽象话语MDP*，我们现在可以直接从模拟中学习政策。我们使用具有 *经验重放的Q学习*来学习该策略，因为它很简单，并且被证明是有效的与神经网络参数化策略（Mnih等人，2013，Lin 1993）。对于体验重播，我们使用大小为1000 的内存缓冲区。我们使用了一种-greedy的勘探方案。我们用贴现因子实验*γ*＆Element; {0 *。*1 *，* 0 *。*2 *，* 0 *。*5}。如前所述，使用Adam更新参数。为了降低过度配置的风险，我们只使用Q-learning来训练与最终输出层相关的权重和跳过连接（如图2中虚线所示）。

训练分两个阶段进行。我们训练100 集的政策。然后，我们评估100 集的平均回报率的政策 。之后，我们继续训练另外100 集的策略。在评估过程中，每个对话历史都从一组对话历史*H*Eval中抽取出来，*H*Eval与训练时使用的*H*训练集不对。这样可以确保政策不会*过分适应*我们有限的对话历史。对于每个超参数组合，我们在400 到600 之间的策略发作。我们选择执行最佳平均回报率的策略。为了保持符号简介，我们将此政策称为*Q学习AMT*。

     4.8        初步评估

在本节中，我们对响应模式选择政策进行了初步评估。

表3：非政府评估预期（平均）Alexa用户分数和测试集上的时间步长数（不包括优先级响应）。

|  |  |  |
| --- | --- | --- |
| 政策 | Alexa用户得分 | 时间步 |
| *监督AMT* | 2 *。*06 | 8 *。*19 |
| *监督学习奖励* | 0 *。*94 | 3 *。*66 |
| *非政策性强化* | **2***。***45** | **10***。***08** |
| *非政策性强化学习奖励* | 1 *。*29 | 5 *。*02 |
| *Q学习AMT* | 2 *。*08 | 8 *。*28 |

非政策评估：评估选择政策的一种方法是使用等式中给出的非政策评估。（15）。该方程式提供了每个政策下预期的Alexa用户分数的估计。[[22]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn22" \o ")如前所述，相同的方程可以用于估计的时间步长的预期数每集（不包括优先级响应）。

预期（平均）Alexa用户分数和每集的时间步长数（不包括优先级响应）在表3中给出。在这里我们观察到，*Off-policy REINFORCE*表现最好，之后是*Q-learning AMT*和*Supervised AMT* wrt预期的Alexa用户得分了。*离职政策强化*达到2 *。*45岁，主要17岁*。*比第二好表现型*Q-learning AMT*提高8％ 。然而，这个优点应该是用一粒盐。如前所述，方程式中的非政府评估 （15）是一个有偏估计，因为重要性权重被截断。此外，*非政策性强化措施*已经受到专门培训，以最大限度地提高这一偏倚的估计量，而所有其他政策都经过培训，以最大化其他目标职能。类似地，wrt预期的时间步长数，*Off-Policy REINFORCE*达到了*Q-learning AMT*和*Supervised AMT*之后的最高时间步长。*像以前一样，我们应该把这个结果用一粒盐，因为这个评价也是有偏见的，没有考虑到优先反应。此外，不清楚的是增加时间步长将增加用户分数。*然而，*非政策性强化*，*Q学习AMT*和*监督AMT*似乎是我们进一步实验的主要候选人。

响应模式选择频率：图8显示了*监控AMT*，

*非政策性增强*和*Q-learning AMT*选择不同的响应模型。*我们观察到*

图8：AMT标签测试数据集上的受*监督AMT*，*非政策增强*和*Q学习AMT的*响应模型的响应模型选择概率。

使用*Off-policy REINFORCE*学习*的政策*倾向于更倾向于使用*Alicebot*对其他模式的响应。该*Alicebot*反应是在系统中，这表明最安全，最主题的依赖，通用响应中*关政策巩固*已学到了高度*厌恶风险的策略*。另一方面，*Q学习AMT*政策选择*Alicebot的*反应，*这些*反应比*Off-Policy REINFORCE*和*Supervised AMT*都要少得多。相反，*Q学习AMT*倾向于倾向于从华盛顿邮报和Google搜索结果中提取的响应。这些反应在语义上更丰富，并且有可能使用户更深入地参与特定的主题，但是它们也更有风险（例如，一个坏的选择可能会破坏整个对话）。这表明*Q-learning AMT*已经学会了一个更具*风险的策略*。这种差异的一个可能的解释是使用模拟训练*Q学习AMT*。通过从模拟学习在线，该策略已经能够探索新的行动，并发现持续多个时间步骤的高级策略。特别是，该政策已被允许尝试采取更高风险的行动，并学习*修复*或*落后策略*，以处理危险行为失败的情况。这也可以解释其对*BoWFactGenerator*响应的更强烈的偏好，这可能是通过输出有关当前主题的事实陈述来作为回退策略。这对于*非政策性强项来说*很难学习，因为在数据中稀疏地观察到这种高级别策略的行为顺序，并且在被观察时，相应的回报（Alexa用户分数）具有高方差。

第二个观察结果是，*Q-learning AMT*对于*Initiatorbot*在三个策略中最为偏爱。这可能表明，*Q学习AMT*倾向于*系统主动策略*（例如系统尝试通过提问，改变主题等来保持对话控制的策略）。需要进一步分析来确认这一点。

摘要话语MDP评估接下来，我们可以评估每个策略WRT模拟的性能*摘要话语MDP*。我们模拟每个政策下的500集，并评估其平均回报率，每个时间步长的平均回报和对话长​​度。除了评估前面描述的五项政策，我们也评估3种启发式策略：1）随机称为策略选择响应*随机*，2）政策只选择*Alicebot*反应称为*Alicebot*，[[23]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn23" \o ") ; 3）政策选择*Evibot*回应在可能的情况下，*Alicebot*反应否则称为*Evibot + Alicebot*[[24]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn24" \o ")。评估这些模型将有助于验证近似MDP。

表4：使用*抽象话语MDP* wrt平均回报，每个时间步长的平均回报和开发人员的平均发作长度（± 标准偏差）的政策评估。奖励功能基于*监督AMT*。

|  |  |  |  |
| --- | --- | --- | --- |
| 政策 | 平均回报 | 每个时间段的平均报酬 | 平均对话长度 |
| *随机* | -32 *。*18±31 *。*77 | -0 *。*87±0 *。*24 | 34 *。*29±33 *。*02 |
| *Alicebot* | -15 *。*56±15 *。*61 | -0 *。*37±0 *。*16 | 42 *。*01±42 *。*00 |
| *Evibot + Alicebot* | -11 *。*33±12 *。*43 | -0 *。*29±0 *。*19 | 37 *。*5±38 *。*69 |
| *监督AMT* | - **6***。***46** ± **8***。***01** | - **0***。***15** ± **0***。***16** | **42***。***84** ± **42***。***92** |
| *监督学习奖励* | -24 *。*19±23 *。*三十 | -0 *。*73±0 *。*27 | 31 *。*91±30 *。*09 |
| *非政策性强化* | - **7***。***30** ± **8***。***90** | - **0***。***16** ± **0***。***16** | **43***。***24** ± **43***。***58** |
| *非政策性强化学习奖励* | -10 *。*19±11 *。*15 | -0 *。*28±0 *。*19 | 35 *。*51±35 *。*05 |
| *Q学习AMT* | - **6***。***54** ± **8***。***02** | - **0***。***15** ± **0***。***18** | **40***。***68** ± **39***。***13** |

结果如表4所示。我们观察到，*监督AMT*在每个时间步长方面表现最好的平均回报率和平均回报。但是，这并不奇怪。MDP中的奖励功能定义为*监督AMT*，因此通过构建，该策略实现了每个时间步长的最高回报。接下来我们观察到*Q学习AMT*与*监督AMT是一致的*，都达到了-0 *。*每个时间步15次平均报酬。排在第二位的是*Off-Policy ReINFORCE*，每个时间段的平均回报为-0 *。*16 。然而，*非政策性强力*也达到最高平均对话长度43 *。*24 。在频谱的另一端，如预期的那样，*随机*政策执行最差的所有指标。相比之下，*Alicebot*和*Evibot + Alicebot都*能够更好地执行所有指标，*Evibot + Alicebot*在三个启发式策略中实现了每个时间步长的最佳平均回报率和平均回报。这将验证*抽象话语MDP*作为培训和评估政策的环境。总的来说，*离职政策*

*强化*，*Q学习AMT*和受*监督的AMT*在初步评估中仍然是表现最好的模式。

图9：应急表对比*监督AMT*和

*Q学习AMT*。矩阵中的单元格显示了*监督AMT*策略选择行响应模型的次数，并且*Q学习AMT*策略选择列响应模型。通过在*抽象话语MDP中*的Q学习策略下模拟500集，计算细胞频率。请注意，从Reddit中检索回应的所有模型已经聚集到*Reddit*类*模型中*。

最后，我们将*Q学习AMT*与*Supervised AMT*进行比较，说明了在*抽象话语MDP中*模拟的情节中状态。如图9所示，这两个策略分散了几个响应模型。当*监督AMT*将选择与主题无关的通用*Alicebot*和 *Elizabot*响应时，*Q学习AMT*通常选择*BoWFactGenerator*，*Initiatorbot*和*VHREDWashingtonPost*响应。例如，有347个*监控AMT*选择*Alicebot的*实例，但是在*Q学习AMT中*选定的*BoWFactGenerator*。同样地，在哪里

*受监督的AMT*将具有优先的通用*VHREDSubtitle*响应，*Q学习AMT*通常选择来自*BoWFactGenerator*，*InitiatorBot*和*VHREDRedditSports的响应*。这支持我们以前的分析，表明 *Q-learning AMT*已经学习了一个更具有风险的策略，其中涉及具有语义丰富内容的响应模型。

在下一节中，我们用真实用户评估这些策略。

5        A / B测试实验

为了评估上一节中描述的对话管理员策略，我们进行A / B测试实验。

在每个A / B测试实验期间，我们评估选择响应模型的几个策略。当Alexa用户开始与系统的对话时，它们被自动分配到一个随机策略，然后记录他们的对话和最终分数。然而，Alexa用户的分布势必随着时间的推移而改变。不同类型的用户将根据一天中的时间，平日和节日季节来使用系统。此外，用户对我们系统的期望随着时间的推移而改变，因为它们与竞争对手中的其他社交媒体进行交互。换句话说，我们必须将Alexa用户分布视为非平稳随机过程。因此，表5：使用六种不同策略进行的第一次A / B测试实验（± 95％置信区间）。星\*表明政策明显优于其他95％统计显着水平的政策。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 政策 | 用户得分 | 对话长度 | 波什。话语 | 负。话语 |
| *Evibot + Alicebot* | 2 *。*86±0 *。*22 | 31 *。*84±6 *。*02 | 2 *。*80％±0 *。*79 | 5 *。*63％±1 *。*27 |
| *监督AMT* | 2 *。*80±0 *。*21 | 34 *。*94±8 *。*07 | **4***。***00**％± **1***。***05** | 8 *。*06％±1 *。*38 |
| *监督学习奖励* | 2 *。*74±0 *。*21 | 27 *。*83±5 *。*05 | 2 *。*56％±0 *。*70 | 6 *。*46％±1 *。*29 |
| *非政策性强化* | 2 *。*86±0 *。*21 | **37***。***51** ± **7***。***21** | 3 *。*98％±0 *。*80 | *。*25％±1 *。*28 |
| *非政策性强化学习奖励* | 2 *。*84±0 *。*23 | 34 *。*56±11 *。*55 | 2 *。*79％±0 *。*76 | 6 *。*90％±1 *。*45 |
| *Q学习AMT* \* | **3***。***15** ± **0***。***20** | 30 *。*26±4 *。*64 | 3 *。*75％±0 *。*93 | **5***。***41**％± **1***。***16** |

我们采取两个步骤来减少用户之间的混淆因素和相关性。首先，在每个A / B测试实验期间，我们同时对所有感兴趣的政策进行评估。这确保了我们与大多数相同数量的用户与每天和工作日的每个策略wrt时间进行交互。这最大限度地减少用户分配的变化对该时段*内*最终用户分数的影响。然而，由于A / B测试实验之间的用户分布的变化，我们仍然无法比较政策性能 *跨越*A / B测试实验。其次，我们丢弃来自返回用户（即已经对系统进行过一次评估的用户）的分数。返回系统的用户可能会受到以前与系统的交互的影响。例如，以前对系统有积极经验的用户可能偏向于在下一次互动中获得高分。此外，返回系统的用户更可能属于与仅尝试系统一次的用户不同的用户群体。这组用户本来可以拥有更多的空闲时间，并且比其他用户更愿意参与社交活动。放弃返回用户分数可确保评估不会偏向用户的子群体。通过丢弃回复用户的分数，我们还确保评估每个用户一次。最后，应该注意的是，我们忽略了Alexa用户没有给出分数的对话。这不可避免地偏向于我们的评估，因为不提供分数的用户可能对系统不满意或者期待着不同的功能（例如*非会话活动*，如播放音乐，玩游戏或参加测验）。*一个潜在的补救办法是让所有对话由第三方进行评估（例如，通过在Amazon Mechanical Turk上提出人为注释来评估对话），但这超出了这项工作的范围。*

     5.1       A / B测试实验＃1

2017年7月29日至2017年8月6日进行了第一次A / B测试实验。

我们测试了六个对话经理政策：*Evibot + Alicebot*，受*监督的AMT*，受*监督的学习奖励*，*非政策性强化*， *非政策性强化学习奖励*和 *Q-learning AMT*。对于*非政策性增强*和*非政策性增强学习奖励*，我们使用等式中定义的贪心变量。（5）。

此次实验发生在亚马逊Alexa奖比赛早期。这意味着Alexa用户对我们的系统的期望很少（例如，系统可以对特定主题进行交涉的期望，或系统可以进行*非会话活动*，例如玩文字游戏或参加测验）。此外，7月29日至8月6日这个时期与美国的暑假重叠。这意味着我们可能会期待更多的孩子与系统相互作用，而不是在其他季节。

政策评估结果见表5. [[25]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn25" \o ")[[26]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn26" \o ")该表显示了平均Alexa用户分数，平均对话长度，正用户话语的平均百分比和负面用户话语的平均百分比。总共在收回用户后收集了一千多个用户评分。[[27]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn27" \o ")每个政策都由大约二百名独特的Alexa用户评估。

正如我们初步评估的预期，我们观察到*Q学习AMT*和*Off-policy*

*加固*所有政策WRT用户分数中表现最佳。 *Q学习AMT*的平均用户得分为3 *。*15 ，这显着高于95％统计学显着水平的所有其他政策，采用单尾双样本t检验。有趣的是，*非政策性强化*实现了最长的对话，平均为37次*。*版本02 */* 2 = 18 *。*每对话51 次。这表明*Offpolicy REINFORCE*产生了高度互动和互动的对话。然而，*非政策性强力*与*Q学习AMT*相比，用户话语的百分比略有增加，情绪负面。这可能表明更长的对话还包括一些沮丧的交互（例如重复同样的问题或陈述的用户，希望系统将在下一次返回更有趣的响应）。其余政策在2 之间实现了Alexa的平均用户分数*。*74 和2 *。*86 ，用启发式政策*Evibot + Alicebot*获得2 *。*86 。这表明其他政策还没有学习如何比*Evibot + Alicebot*启发式更合适地选择答案。

总之，结果表明，*风险容忍*所学到*Q-AMT的学习*策略的所有策略中表现最好。这表明通过*抽象话语MDP*中的模拟来学习政策 可能是开发开放域名社交网站的有效途径。此外，*离职*政策强化的表现表明，直接向Alexa用户分数优化政策也可能产生改善。但是，需要进一步调查。

Initiatorbot评估本实验还允许我们分析由*Initiatorbot*给出的不同对话启动短语的结果。我们通过计算40个可能短语中的每一个的平均Alexa用户分数来进行此分析。我们发现与消息相关的短语（例如*“你跟随消息”*），政治（例如*“你想谈论政治吗？”*）和旅行（例如*“告诉我，你喜欢去哪里度假？ “*）在所有政策中表现不佳。另一方面，与动物相关的短语（例如*“你有宠物吗？”*和*“你可以想到的最可爱的动物”*），电影（例如*“我们来谈谈电影，你看过的最后一部电影是什么？”*）和食物（例如*“让我们谈谈食物，你最喜欢的食物是什么？”*）在所有政策中表现良好。例如，*Initiatorbot*询问有关新闻和政治问题的对话的平均Alexa用户得分只有2 *。*前两个系统的91 （*Off-Policy REINFORCE*和 *Q-learning AMT*）。意思是说，*Initiatorbot*询问有关动物，食物和电影的对话，相应的平均Alexa用户得分为3 *。*17。我们预计会话话题会影响用户参与，但令人惊讶的是，这些特定主题（动物，食物和电影）是最受欢迎的主题。一个可能的解释是，我们的系统在新闻，政治和旅游话题方面表现不佳。然而，该系统已经有几个专门讨论新闻和政治的响应模型：六个序列到序列模型提取了来自Reddit新闻和Reddit政治的响应，两个模型提取了华盛顿邮报用户评论的反应和*BoWTrump*模型提取Donald J特朗普的Twitter个人资料。此外，*Evibot*还能够回答许多有关新闻政治的事实问题

*BoWFactGenerator*包含了与消息和政治有关的数百个事实。*因此，用户对动物，电影和食物等主题的喜好可能会更为合理。一个可能的解释是用户的年龄组。在检查我们的日志时，我们观察到许多孩子正在与我们的系统进行互动，这些孩子们更喜欢谈论动物，电影和食物，而不是消息，政治和旅游，这并不奇怪。*

     5.2       A / B测试实验＃2

第二次A / B测试实验是在2017年8月6日至2017年8月13日期间进行的。我们测试了两项对话经理政策：*非政策性增强*和 *Q-学习AMT*。如前所述，我们使用等式中定义的*Off-policy REINFORCE*的贪心变体。（5）。

该实验发生在亚马逊Alexa奖比赛结束时。这意味着很多

Alexa用户已经在竞争中与其他社交媒体进行了互动，因此可能对系统产生了期望。这些期望可能涉及对特定主题*的交谈*或从事*非对话活动*，如玩游戏。此外，8月6日至8月13日这个时期与美国暑假结束和学年开始重叠。这意味着我们应该期望与以前的A / B测试实验相比，较少的孩子与系统进行交互。

表6：具有两种不同策略（± 95％置信区间）的第二次A / B测试实验。明星\*表示政策明显优于其他95％统计显着水平的政策。

|  |  |  |
| --- | --- | --- |
| 政策 | 用户评分              对话长度 | 波什。说话       阴 话语 |
| *离职政策强力* \* | **3***。***10** ± **0***。***12 34***。***86** ± **3***。***66** | 3 *。*21％±0 *。*44             7 *。*96％±0 *。*82 |
| *Q学习AMT* | 2 *。*93±0 *。*12           32 *。*06±3 *。*58 | **3***。***42**％± **0***。***48 7***。***56**％± **0***。***81** |

政策评估结果如表6所示。总共在收回用户后收集了约八百个用户评级。因此，每个政策由大约四百名独特的Alexa用户评估。

我们观察到，*非政策性增强*和*Q-学习AMT的*表现优于前面实验中的政策。然而，在本实验中，*Off-Policy REINFORCE*平均Alexa用户得分为3 *。*06 而*Q学习AMT*的得分较低只有2分*。*95 。尽管如此，*非政策*性增长在统计学上没有明显好转。此外，*离职政策强化*实现了负面用户话语的百分比略高，并且比较了*Q学习AMT*的正面用户话语的百分比略低。

如前所述，与以前的A / B测试实验相比，性能差异可能是由于用户配置文件和用户期望的变化。在这个时候，更多的Alexa用户已经与其他团队的社交媒体进行了互动。同时，所有的社交网站都在不断发展。因此，用户对我们系统的期望可能会更高。此外，由于暑假已经结束，预计少儿和更多的成年人将与我们的系统进行互动。这些成年人对系统有更高的期望似乎是合理的，甚至更有可能他们不那么有趣，对错误的容忍也较少。鉴于这种变化在用户配置文件和预期，*风险容忍策略*所学到*Q学习AMT*与*离职政策强化*的*风险厌恶*策略相比，政策可能不佳。

     5.3      讨论

这两个政策*Q-learning AMT*和*Off-Policy REINFORCE*已经显示出比所有其他政策的重大改进，包括使用监督学习和启发式政策学习的政策。如前所述，*Q-learning AMT*政策在竞争半决赛中获得了与一些顶级球队相当的Alexa用户平均​​水平。意味着，*脱离政策的强力*在对话中转弯的次数很多，这表明由此产生的对话交互性更强。结果证明了整体综合方法的优点，其中许多不同的模型产生自然语言响应，系统策略选择一个响应。结果还强调了使用强化学习技术学习政策的优势。通过优化使用*抽象话语MDP*中的真实用户分数或奖励来实现最大化的策略，具有代理奖励功能，我们已经证明可以实现显着的收益，既可以实现现实用户分数和对话转数。

6        未来工作

     6.1       个性化

未来研究的一个重要方向是个性化，即建立每个用户个性，意见和兴趣的模型。这将允许系统通过使响应模型适应用户的已知属性来提供更好的用户体验。我们正在实现一个给定用户ID的状态机，从数据库检索用户的相关信息属性。如果缺少特定的用户属性，则状态机将向用户询问相关信息并将其存储在数据库中。一个重要的用户属性是用户的名字。如果在数据库中没有找到名称，状态机可能会询问用户要调用什么，然后从用户的响应中提取名称。如果检测到个人姓名，它存储在数据库中以供其他模块可用以插入其响应。名称检测进行如下。首先，我们将响应与一小段模板匹配，例如“我的名字是...”或“打电话给我...”。然后，我们使用结果匹配的部分语音（POS）标签来检测名称的结束边界。为了避免由于错误的POS标签而过早地修改名称，我们还将字词与1990年美国普查数据中的常用名称列表进行了匹配[[28]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftn28" \o ")。

将来，我们计划从以前与每个用户的互动中探索学习用户嵌入。这将允许系统变得更加个性化，通过提供我们的反应模型与超出即时对话历史的附加背景。

7        结论

我们为亚马逊提出了一个新的大型综合型对话系统框架

Alexa奖比赛。我们的系统利用各种机器学习技术，包括深入学习和强化学习。我们开发了一套新的自然语言检索和生成深入学习模型，包括循证神经网络，序列序列模型和潜变量模型。此外，我们开发了一种新颖的强化学习程序，并对现有用户的A / B测试实验中现有的强化学习方法进行了评估。这些创新使我们能够对我们的基准系统进行实质性改进。在一个规模1 - 5 ，我们表现最好的系统达到了平均用户评分3 *。*15手工制作国家和规则的数量很少，而且不进行*非会话活动*（如玩游戏）。这与半决赛中的一些顶级系统相当。此外，同一系统平均高14 *。*5 - 16 *。*每次会话0 次，这表明我们的系统是比赛中最具*互动性*和*吸引力的*系统之一。由于我们几乎所有的系统组件都是可训练的机器学习模型，所以随着更多的互动和额外的数据，系统可能会大大提高。

致谢

感谢Aaron Courville，Michael Noseworthy，Nicolas Angelard-Gontier，Ryan Lowe，Prasanna Parthasarathi和Peter Henderson在整个Alexa奖比赛中对系统架构，众包和强化学习提供了有用的建议。我们感谢Christian Droulers为基于文本的聊天构建图形用户界面。我们感谢亚马逊通过Amazon Web Services平台提供特斯拉K80 GPU。NVIDIA公司捐赠了一些用于本研究的Titan X GPU。作者承认NSERC，加拿大研究主席，CIFAR，IBM研究，Nuance基金会，Microsoft Maluuba和Druide Informatique Inc.资助。

参考

Ameixa，D.，Coheur，L.，Fialho，P.＆Quaresma，P.（2014），Luke，我是你的父亲：通过使用电影字幕处理域外请求，*在* “智能虚拟代理”斯普林格。

Aust，H.，Oerder，M.，Seide，F.＆Steinbiss，V.（1995），“The philips automatic train timetable information system”，*Speech Communication*17（3）。

Bird，S.，Klein，E.＆Loper，E.（2009），*Natural Language Processing with Python*，O'Reilly Media。

Blunsom，P.，Grefenstette，E.＆Kalchbrenner，N。（2014），A convolutional neural network for modeling sentence，*in* “Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics”，Proceedings of the 52nd Annual Meeting of the计算协会

语言学。

Bohus，D.，Raux，A.，Harris，TK，Eskenazi，M.＆Rudnicky，AI（2007），Olympus：a open-source framework for conversational spoken language interface research，*in* “Proceedings of the workshop on bridging the gap ：Academic and industrial research in dialog technology“，计算语言学协会，第32-39页。

Breiman，L.（1996），“Bagging predictors”， *Machine learning* 24（2），123-140。

Charras，F.，Duplessis，GD，Letard，V.，Ligozat，A.-L. ＆罗塞特，S.（2016），开放域和休闲会话代理，比较systemresponse检索模型*在* “上聊天机器人和会话代理技术研讨会”。

Cho，K.，van Merrienboer，B.，Gulcehre，C.，Bahdanau，D.，Bougares，F.，Schwenk，H.＆Bengio，Y。（2014），Learning phrase representation using rnn encoder-decoder for statistical machine 翻译，*在* “EMNLP”。

Colby，KM（1981），“建模偏执*主义* ”，*行为与脑科学* 4。

Ferrucci，D.，Brown，E.，Chu-Carroll，J.，Fan，J.，Gondek，D.，Kalyanpur，AA，Lally，A.，Murdock，

JW，Nyberg，E.，Prager，J.et al。（2010），“建筑沃特森：深层项目概述”，*AI杂志* 31（3）。

Glorot，X.，Bordes，A.＆Bengio，Y.（2011），Deep sparse rectifier neural networks，*in* “Proceedings of the Fourthteenth International Conference on Artificial Intelligence and Statistics”，第315-323页。

Im，J.（2017）。

网址：*http* : *//search.aifounded.com/*

JurcíCEK，F.，Dušek，澳，PLATEK，O＆Žilka，L.（2014），亚历克斯：统计对话系统框架，*在* '国际会议上的文字，语音和对话'，施普林格，页。 587-594。

Kingma，D.＆Ba，J.（2015），Adam：A method for stochastic optimization，*in* “ICLR”。

Kingma，DP＆Welling，M.（2014），“Auto-encoding variational Bayes”，*ICLR*。

基罗斯，R.，朱，Y.，Salakhutdinov，RR，泽梅尔，R.，Urtasun，R.，托拉尔瓦，A。＆菲德勒，S。（2015），跳过深思熟虑载体，*在* 'NIPS'。

Koren，Y.，Bell，R.＆Volinsky，C。（2009），“Matrix factorization techniques for recommender systems”，*Computer* 42（8）。

林氏 （1993），使用神经网络的机器人加固学习，技术报告，卡内基梅隆大学匹兹堡PA计算机科学学院。

Liu，C.-W.，Lowe，R.，Serban，IV，Noseworthy，M.，Charlin，L.＆Pineau，J。（2016），“如何评估你的对话系统：无监督评估的经验研究对话响应生成的指标“，*arXiv：1603.08023*。

Lowe，R.，Noseworthy，M.，Serban，IV，Angelard-Gontier，N.，Bengio，Y.＆Pineau，J.（2017），Towards a automatic Turing test：Learning to evaluate dialog responses，*in* “ACL” 。

Lowe，R.，Pow，N.，Serban，I.＆Pineau，J。（2015），The Ubuntu Dialogue Corpus：A Large Dataset for Research in Nonstructured Multi-Turn Dialogue Systems，*in* “SIGDIAL”。

*Marelli，M.，Bentivogli，L.，Baroni，M.，Bernardi，R.，Menini，S.＆Zamparelli，R。（2014），Semeval2014 task 1：Evaluation of compositional distributional semantic models on full sentence through semantic relatedness and 在* “SemEval Workshop，COLING”中的文字内容。

麦格拉斯汉，S.，弗雷泽，N.，吉尔伯特，N.，Bilange，E.，Heisterkamp，P。＆你几乎，N。（1992），对话管理电话信息系统，*在* 'ANLC'。

Mikolov，T.，Sutskever，一，陈，K.，科拉多，GS＆院长，J.（2013），分布式的单词和短语以及它们的组合性的陈述，*在* 'NIPS'。

Mnih，V.，Kavukcuoglu，K.，Silver，D.，Graves，A.，Antonoglou，I.，Wierstra，D.＆Riedmiller，M。（2013），“使用深加强学习玩atari”，*arXiv preprint arXiv ：1312.5602*。

奈尔，V＆韩丁，GE（2010），整流后的线性单位完善受限玻尔兹曼机，*在*第807-814“机器上的第27届国际会议学习（ICML-10）论文集”。

Ng，AY，Harada，D.＆Russell，S。（1999），Policy invariance under reward transformations：Theory and application to reward shaping，*in*“ICML”，Vol。99，pp。278-287。

Nguyen，T.，Rosenberg，M.，Song，X.，Gao，J.，Tiwary，S.，Majumder，R.＆Deng，L。（2016），“MS MARCO：A Human Generated MAchine Reading COmprehension Dataset” ，*arXiv preprint arXiv：1611.09268*。

潘宁顿，J.，Socher，R。＆曼宁，CD（2014），手套：对于单词表示全局矢量，*在* 'EMNLP'，卷。14。

Precup，D.（2000），“非政策性政策评估的资格追踪”，*计算机科学系教师出版系列*。

Precup，D.，萨顿，RS和达斯古普塔，S.（2001），离政策的时间差与函数逼近学习，*在* 'ICML'。

Rezende，DJ，Mohamed，S.＆Wierstra，D。（2014），Stochastic backpropagation and approximation in deep generative models，*in* “ICML”，pp。1278-1286。

Serban，IV，Sordoni，A.，Lowe，R.，Charlin，L.，Pineau，J.，Courville，A.＆Bengio，Y。（2017），A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues，*in* 'AAAI'。

Shawar，BA＆Atwell，E.（2007），Chatbots：他们真的有用吗？*在* “LDV论坛” 22。

Simpson，A.＆Eraser，NM（1993），黑盒子和玻璃盒评估日ial系统， *在* “第三届欧洲语音通信和技术会议”。

Socher，R.，Perelygin，A.，Wu，JY，Chuang，J.，Manning，CD，Ng，AY，Potts，C。（2013年）为在情绪树库语义组合性，递归深模型，*在*卷“在自然语言处理（EMNLP）的实证方法研讨会论文集”。1631，p。1642。

Stolcke，A.，Ries，K.，Coccaro，N.，Shriberg，E.，Bates，R.，Jurafsky，D.，Taylor，P.，Martin，R.，Van Ess-Dykema，C.＆Meteer， M.（2000），“Dialogue act modeling for automatic tagging and recognition of conversational speech”，*Computational linguistics* 26（3）。

Stone，B.＆Soper，S.（2014），“ *Bloomberg LP* ”，“亚马逊推出一款听力，谈话，音乐演奏家” 。检索2014-11-07。

Suendermann-Oeft，D.，Ramanarayanan，V.，Teckenbrock，M.，Neutatz，F.＆Schmidt，D。（2015），Halef：An open-source standard-compliant telephony-based modular spoken dialog system：A review and 一个展望，*在* “自然语言对话系统和智能助手”，Springer。

萨顿，RS＆巴托，AG（1998），*强化学习：简介*，数字1 *在*'1'，麻省理工学院出版社剑桥。

Wallace，RS（2009），“解剖学的爱丽丝”，*解析图灵测试*。

Weizenbaum，J.（1966），“Eliza-a computer program for the study of natural language communication between man and machine”，*ACM*9（1）。

威廉姆斯，JD（2011年），统计对话系统的公共使用的实证分析，*在*

“Proceedings of the SIGDIAL 2011 Conference”，计算语言学协会，第130-141页。

Williams，JD，Raux，A.＆Henderson，M。（2016），“Introduction to the special issue on dialog state tracking”，*Dialogue＆Dialogue* 7（3），1-3。

威廉斯，J.，RAUX，A.，德兰，D.＆黑，A.（2013），对话状态跟踪的挑战，*在* 'SIGDIAL'，第404-413。

Williams，RJ（1992），“Simple statistics gradient-following algorithms for connectionist reinforcement learning”，*Machine learning*8（3-4）。

Wu，Y.，Schuster，M.，Chen，Z.，Le，QV，Norouzi，M.，Macherey，W.，Krikun，M.，Cao，Y.，Gao，Q.，Macherey， 。（2016），“Google的神经机器翻译系统：弥合人与机器翻译之间的差距”，*ArXiv预印版：1609.08144*。

Yu，L.，Hermann，KM，Blunsom，P.＆Pulman，S.（2014），Deep learning for answer sentence selection， *in* “NIPS，Workshop on Deep Learning”。

Yu，Z.，Xu，Z.，Black，AW＆Rudnicky，AI（2016），Strategy and policy learning for non-taskoriented conversational systems。，*in*“SIGDIAL”。

Zhu，Y.，Kiros，R.，Zemel，R.，Salakhutdinov，R.，Urtasun，R.，Torralba，A.＆Fidler，S。（2015），Aligning books and movies：Towards story-like visual explanation 看电影和看书，*在* “ICCV”。

[[1]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref1" \o ")在整个半决赛中，我们进行了几次A / B测试实验，以评估我们系统的不同变体（参见第5节）。得分3*。*15 是基于在2017年7月29日至8月6日期间表现最好的系统。得分不是基于平均值系统的所有变体的得分的排行榜（包括监督学习系统和启发式基准系统）。

[[2]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref2" \o ")竞赛规则禁止我们说明任何其他团队的表现。

[[3]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref3" \o ")如果存在多个*优先级*响应，则模型的排序决定了返回哪个响应。

[[4]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref4" \o ")我们使用以下实现：[https](https://gist.github.com/bebraw/273706)：[//gist.github.com/bebraw/273706。](https://gist.github.com/bebraw/273706)

[[5]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref5" \o ")告诉故事的请求可能是用户与其他团队的漫游器互动的副作用，通常会强调*非会话活动*，例如讲故事，玩测验和文字游戏。

[[6]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref6" \o ")参见[www.omdbapi.com。](http://www.omdbapi.com/)[这不应该与IMDB混淆。](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftnref6)

[[7]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref7" \o ")我们使用维基百科2014 + Gigaword 5：[https](https://nlp.stanford.edu/projects/glove/)://nlp.stanford.edu/[projects / glove /](https://nlp.stanford.edu/projects/glove/)上训练的手套嵌入 [。](https://nlp.stanford.edu/projects/glove/)

[[8]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref8" \o ")对于*VHREDWashingtonPost*，基于当前对话和新闻文章关键词之间的余弦相似性提取*K个*响应。*K*取决于在一定的余弦相似性阈值之上的一组新闻文章中的用户评论的数量。

[[9]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref9" \o ")对于*VHREDSubtitles*，基于每个单词的一个热向量来计算余弦相似度。

[[10]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref10" \o ")触发短语可能有多个响应。[在这种情况下，随机选择一个响应。](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftnref10)

[[11]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref11" \o ")一些触发短语没有关键字。[在这种情况下，匹配仅基于语义相关性。](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftnref11)

[[12]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref12" \o ")我们使用预先训练的Word2Vec嵌入：[https](https://code.google.com/archive/p/word2vec/):[//code.google.com/archive/p/word2vec/。](https://code.google.com/archive/p/word2vec/)

[[13]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref13" \o ")这个模式是在比赛结束后实施的，但是这里包含了完整性。

[[14]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref14" \o ")通常，由于一些响应模型仅适用于特定的用户话语输出响应，候选响应的数量还依赖于状态而变化。

[[15]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref15" \o ")我们使用预先训练的Word2Vec嵌入：[https](https://code.google.com/archive/p/word2vec/):[//code.google.com/archive/p/word2vec/。](https://code.google.com/archive/p/word2vec/)

[[16]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref16" \o ")随机采样对我们的目标是有利的，因为它确保候选人对频繁的用户陈述和问题的响应往往被更多的土耳其人注释。[这增加了这种话语的平均注释精度，这反过来又增加了评分模型对于这种话语的准确性。](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftnref16)

[[17]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref17" \o ")所述的掩蔽是不完美的。[因此，我们还指示土耳其人的任务可能包含亵渎和淫秽语言。此外，还应该注意的是，亚马逊机械工程师只雇用成年人。](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftnref17)

[[18]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref18" \o ")通过忽略没有Alexa用户分数的对话，我们在我们的奖励模型中引入了显着的偏见。特别是，没有提供分数的用户似乎很可能发现系统很差，或者缺乏他们预期的特定功能/功能（例如*非会话活动*，如玩游戏或参加测验）。[医疗统计中出现相关问题，患者接受治疗后，未发现其结果。](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftnref18)

[[19]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref19" \o ")这通过手动检查会话日志来确认，大多数会话都有多个语音识别错误。[在具有过多语音识别错误（可能是由于嘈杂环境）的对话中，用户的话语清楚地表明对系统的沮丧。](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftnref19)

[[20]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref20" \o ")然而，由于重要权重被削减，*离职后*强化培训程序仍然存在偏见。

[[21]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref21" \o ")例如，如果我们要使用高斯分布，我们必须至少还要指定方差参数。

[[22]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref22" \o ")对于参数化作为动作值函数的策略，我们转换等式 [（2）至](file:///D:\BaiduNetdiskDownload\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\mailabot\milabotweb.htm#_ftnref22)（4）通过设置*˚Fθ*=*Qθ*和装配温度参数*λ*对*离政策加固*发展集。

[[23]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref23" \o ")当没有有效的*Alicebot*响应时，此策略随机选择一个响应。

[[24]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref24" \o ")当没有有效的*Evibot*或*Alicebot*响应时，此策略随机选择一个响应。

[[25]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref25" \o ")

[[26]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref26" \o ")％置信区间的是每个策略的用户的Alexa分数从与它自己的均值和方差的正态分布得出的假设下计算的。这是一个近似值，因为Alexa的用户得分只对间隔支撑[1 *，*5] 。

[[27]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref27" \o ")在半决赛比赛结束后收集评分，所有评分均由人类注释者转录。

[[28]](file:///D:\\BaiduNetdiskDownload\\%E4%BA%BA%E6%9C%BA%E9%97%AE%E7%AD%94%E7%B3%BB%E7%BB%9FQuestionAnsweringSystem\\mailabot\\milabotweb.htm" \l "_ftnref28" \o ")从：[https](https://deron.meranda.us/data/):[//deron.meranda.us/data/获取。](https://deron.meranda.us/data/)