Dialog-to-Actions：通过动作级生成构建面向任务的对话系统

Yuncheng Hua
Xiangyu Xi*美团集团，中国
devin.hua@monash.edu
xixy10@foxmail.com

Zheng Jiang 美团集团，中国
zjiang@seu.edu.cn

Guanwei Zhang 美团集团，中国
zhangguanwei@meituan.com

Chaobo Sun 美团集团，中国
sunchaobo@meituan.com

Guanglu Wan 美团集团，中国
wanguanglu@meituan.com

Wei Ye 北京大学，中国，中国 wye@pku.edu.cn

arXiv:2304.00884v1 [cs.CL] 3 Apr 2023

摘要

端到端生成方法已在面向任务的对话系统中得到研究和应用。然而，在工业场景中，现有方法面临可靠性（例如，领域不一致的响应、重复问题等）和效率（例如，计算时间长等）方面的瓶颈。本文提出一种基于动作级生成的面向任务的对话系统。具体而言，我们首先从大规模对话中构建对话动作，并将每个自然语言（NL）响应表示为一系列对话动作。随后，我们训练一个序列到序列模型，以对话历史为输入，输出一系列对话动作。生成的对话动作会被转换为口头响应。实验结果表明，我们的轻量级方法取得了具有竞争力的性能，并具有可靠性和效率方面的优势。

CCS 概念

• **计算方法** → **话语、对话与语用学**。

关键词

面向任务的对话系统，动作级生成，Dialog-to-Actions

ACM 引用格式：

Yuncheng Hua、Xiangyu Xi、Zheng Jiang、Guanwei Zhang、Chaobo Sun、GuangluWan 和 WeiYe。2023。Dialog-to-Actions：通过动作级生成构建面向任务的对话系统。载于*ACM Conference'17* 会议论文集。ACM，纽约，纽约州，美国，5 页。https://doi.org/ 10.1145/nnnnnnn.nnnnnnn

*Yuncheng Hua 和 Xiangyu Xi 对本研究的贡献相同。

特此授予许可，允许为个人或课堂使用而制作本作品的全部或部分数字副本或纸质副本，前提是副本不得用于营利或商业利益，且须附上本声明和首页的完整引用。对于本作品中由 ACM 以外其他方拥有版权的部分，必须尊重其版权。允许在注明出处的情况下摘录。若要以其他方式复制、再版、发布到服务器或重新分发至列表，则必须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 申请许可。*Conference'17，2017*年7月，美国华盛顿特区© 2023 计算机协会。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

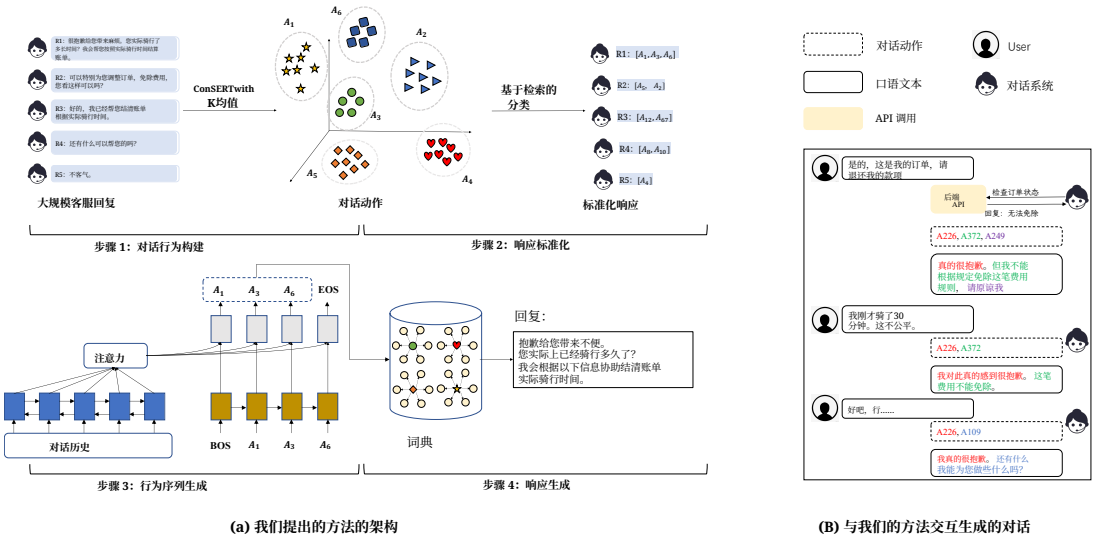
近年来，能够直接生成恰当的自然语言回复或 API 调用的端到端生成式方法，在面向任务的聊天机器人领域得到了深入研究 [2, 5, 9, 16, 22], ，并已被证明对实际业务，尤其是售后客户服务 [1, 7, 8, 13, 14, 19, 21, 24, 25]，具有重要价值。基于大规模预训练语言模型 [10, 11], 的生成式方法具有架构更简单、交互更拟人化的优势。尽管取得了显著进展，我们发现，这些基于 token 级别的生成方法在实际场景中存在以下两项局限。

1. 词元级生成方法的可靠性有限，而可靠性对于工业任务型对话系统至关重要。由于预训练语言模型的特性，模型可能会生成从预训练语料库中学到的回复。在某些情况下，这些回复毫无意义，或与当前业务领域语义不符，从而中断线上交互。更糟的是，模型偶尔会在多轮对话中陷入重复生成回复的状态（例如，反复向用户询问相同的信息）。其他研究人员 [4, 12]和从业者[1]也普遍观察到了上述问题。

2. 词元级生成方法可能无法满足工业系统对效率的要求，尤其是在解码步骤较多时。词元级生成模型的计算时间较长，导致在线对话系统的响应延迟难以接受，尤其是在模型生成的句子长度超过阈值时（例如，如图 3 所示，T5 生成一个 30 个词的句子需要 1,544 毫秒）。由于延迟问题，高峰期可能会有大量服务请求被暂停或阻塞。此外，上述系统所需的计算资源（例如图形处理器）对于小公司而言可能难以负担。

为了解决上述两个问题，本文提出一种基于动作级生成方法的面向任务的对话系统。受 Xi 等人启发， [19], 我们使用**对话动作**表示回复，即一类语义独特且一致、可通过聚类自动获得的回复。Xi 等人 [19] 直接将整个回复视为一个特定的对话动作，而我们将一条回复拆分为多个

[1]https://github.com/microsoft/DialoGPT/issues/45



(a) 我们提出的方法的架构 (B) 与我们的方法交互生成的对话

图 1：系统架构和对话示例。在（b）中，对话动作及其对应的话语片段用相同颜色标记（例如，“A226”和“我真的很抱歉”）。

片段 [6]，每个片段都可以映射到一个对话动作。这样，每条回复都表示为一个对话动作序列。给定对话上下文，使用带有动作级循环解码器的序列到序列模型生成对话动作序列。此外，根据生成的对话动作序列，使用基于频率的采样方法来组合最终回复。由于我们方法的核心组件是生成模型，它以对话上下文为输入，并输出操作，因此我们将该方法命名为 Dialog-To-Actions（简称**DTA**）。与现有基于词元级生成的系统相比，我们的 DTA 具有以下优势：1）可靠性，因为生成的自然语言回复源自预定义的对话动作；2）效率，因为解码空间（即对话动作）和解码步骤都要小得多。

2 框架描述

2.1 概述

我们遵循先前端到端任务型对话系统所采用的工作流程 [2, 9],：系统以对话历史为输入，并生成一个文本字符串，该字符串要么作为对用户的口头员工回复，要么作为 API 调用（例如信息查询、动作执行等）。当调用 API 时，API 返回的信息将被纳入系统的下一条回复中。遵循这一系统交互生命周期的对话示例见图 1 (b)。

我们工作的核心思路是先生成对话动作，再组织成口头响应。为此，我们首先从大规模对话中构建对话动作（步骤 1），并将每个响应表示为一系列对话动作（步骤 2），如图 1 (a) 所示。我们采用带有动作级循环解码器的序列到序列模型来生成对话动作（步骤 3），并进一步利用生成的动作组织成口头响应（步骤 4）。我们以电动自行车租赁业务的售后客户服务为例，用户和工作人员通过短信在线沟通。下面将介绍具体技术细节。

2.2 步骤 1：对话动作构建

对话动作是指一组具有相同语义并表达共同交际意图的话语或话语片段，例如提出请求或查询信息。Xi 等人 [19] 将具有相同语义信息的一组话语视为对话动作，并选择与特定员工动作相对应的回复。然而，这种过度简化的设定，即将整个话语抽象为一个动作，会导致表达能力和可扩展性相对有限。为了使回复更具针对性和灵活性，我们基于员工话语片段而非完整话语构建对话动作。具体而言，我们通过一种基于规则的方法 [6] 将每个话语划分为多个片段。进一步地，遵循 Xi 等人 [19], 的方法，我们采用两阶段方法对这些片段进行聚类。具体而言，利用 ConSERT [20]为每个话语片段生成表示，然后应用 K均值对片段进行聚类。我们通过经验选择聚类数量 K，以平衡纯度和聚类数量，并将每个片段簇视为一个对话动作（例如图 1 (a) 中的 A_1 和 A_2 ）。

2.3 步骤2：响应标准化

响应标准化旨在通过将每个响应映射为一系列对话动作，来规范化（来自大规模对话的）响应。遵循 Yu 等人 [23], 的方法，我们采用一种基于检索的方法，检索与给定输入话语片段最相似的聚类片段，并根据相应的聚类为输入添加标签。如图 2 所示，给定一个输入片段 x，我们使用 BM25 从所有聚类片段中召回排名前 k 的片段 $\{u_1, \ldots\ldots u_k\}$。此外，我们利用一个基于 BERT 的文本相似度计算模型 S 对这些片段 k 重新排序，并选取

与 x相似度最高的片段 $\hat{u}$，记为：

$$\hat{u} = \underset{u_i \in \{u_1,...,u_k\}}{\text{argmax}} \; S(x, u_i) \tag{1}$$

其中，$S(x, u_i)$表示 x与 u_i之间的相似度。随后，为 x标注其所属的对话动作 A_i，$\hat{u}$。

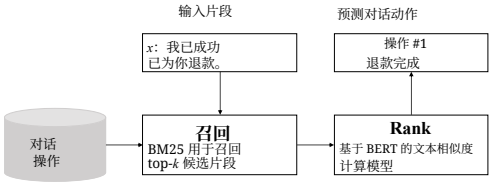


图2：响应标准化流程

此外，我们记录了对话动作与话语片段之间的对应关系，以及话语片段在对话中的出现频率。具体而言，我们使用一个键值字典 $\mathfrak{D}$，其中键表示一个对话动作 A_i，值则是一个嵌套字典，用于记录唯一片段 $\{x_1, \ldots\ldots x_n\}$ 与 A_i 之间的映射关系，以及各片段的出现频率 $\{f_1, \ldots\ldots f_n\}$。该字典 $\mathfrak{D}$ 用于生成口头响应（步骤 4）。

2.4 步骤 3：动作序列预测

给定对话 ID 由用户（U_1）与工作人员（S_i）交替交换的一组话语构成。经过上述精心设计的步骤后，每条工作人员话语 S_i 都表示为一个动作序列 A_{S_i}。在第 m 轮，给定对话历史 $H_m = \{U_{m-w}, S_{m-w}, \ldots\ldots S_{m-1}, U_m\}$，我们提出一种序列到序列模型来生成工作人员响应 S_m。我们首先使用序列到序列模型生成一个动作序列 $A_{S_m} = (A_1^{S_m}, A_2^{S_m}, \ldots\ldots A_k^{S_m})$，其中 k 表示动作序列的长度，然后使用该动作序列生成口头响应（见 § 2.5）。

编码器 给定对话历史 H_m，我们依次拼接 H_m中的所有话语以及每个工作人员话语对应的动作序列 A_{S_i}，形成一个词元序列 $(w_1, \ldots\ldots, w_n)$。我们使用双向长短期记忆网络模型将该词元序列编码为一系列连续表示 $\boldsymbol{H}$：

$$\boldsymbol{H} = (\boldsymbol{h}_1, ..., \boldsymbol{h}_n) = \text{BiLSTM}(w_1, ..., w_n) \tag{2}$$

解码器 考虑到效率要求以及较小的解码空间，我们采用 Luong 注意力方法，并使用长短期记忆网络作为解码器，按如下方式计算时间步 t 的隐藏状态 $\boldsymbol{s_t}$：

$$\boldsymbol{s_t} = \text{LSTM}(\boldsymbol{s_{t-1}}, g(\boldsymbol{y_{t-1}}), \boldsymbol{c_{t-1}}) \tag{3}$$

其中，$\boldsymbol{y_{t-1}}$ 表示步骤 t-1 时对话动作空间上的概率分布，$g(\boldsymbol{y_{t-1}})$表示概率最高的动作。获得隐藏状态 $\boldsymbol{s_t}$ 和上下文向量 $\boldsymbol{c_t}$ 后，我们按如下方式生成时间步 t 的概率分布：

$$\boldsymbol{y_t} = \text{Softmax}(\boldsymbol{W_d}[\boldsymbol{s_t}; \boldsymbol{c_t}]) \tag{4}$$

其中 $\boldsymbol{W_d}$ 是权重参数。给定时间步 t的真实标签 y_t，我们用 $p(y_t|y_{<t}, H_m)$ 表示步骤 t的交叉熵损失，其中 $y_{<t}$表示此前生成的操作。

优化目标定义如下：

$$L_{Gen} = -\sum_{\mathcal{D} \in C} \sum_{\mathcal{H}_m \in \mathcal{D}} \sum_{t=1}^{l_m} p(y_t|y_{<t}, \mathcal{H}_m) \tag{5}$$

其中，C 表示对话集合，l_m 表示对话 D 的第 m轮中对话动作序列的长度。

2.5 步骤4：响应生成

基于 § 2.4 中生成的动作序列，我们为每个动作选取一个话语片段，并依次组合这些片段，从而构成口头响应。考虑到频率较高的片段更可能是工作人员通常使用的规范话语，我们从 $\mathfrak{D}$（在 § 2.3 中构建）中采样片段，并遵循频率越高、片段被选中的可能性越大的原则。通过这种方式，我们确保口头响应的质量和多样性。

3 实验

3.1 实验设置

3.1.1 数据集。我们针对电动自行车租赁业务的中国线上售后客服开展了一项实验。在这一场景中，用户骑行结束后可能忘记锁车，因此会要求客服远程锁车并减少费用。客服需要通过后端API查询订单状态，判断是否可以减免费用。我们从在线服务日志中收集了一周的用户与客服对话。数据统计见表1。对话被随机划分为训练集、开发集和测试集，比例为8:1:1。我们构建了1,420个对话动作（§ 2.2），并将每次API调用视为一个对话动作。该数据集将在线发布。

统计类型	数值
对话	8,363
总轮数	55,576
每段对话平均轮数	6.65
对话动作	1,420

表 1：数据统计。

3.1.2 基线。为了评估我们方法的有效性和效率，我们将其与以下最先进的基线方法进行比较：**(1) 长短期记忆网络**，采用经典的基于 LSTM 的序列到序列架构 [15]；**(2) Transformer**，使用 Transformer [17] 作为编码器和解码器；**(3) CDiaGPT**，一种基于大规模中文对话数据集[18]预训练的 GPT 模型；**(4) T5**，一种使用 CLUE 语料库预训练的文本到文本迁移 Transformer（T5）模型[11]。

3.1.3 评估指标。为了全面评估不同模型的有效性，我们同时进行离线评估（即在数据集上评估）和在线评估（即线上 A/B 测试）。

离线评估 模型以特定的真实对话历史（即上下文）为输入，并生成回复。遵循 Byrne 等人 [2],，我们报告每个模型在测试集上的 BLEU-4 分数。考虑到 API 调用至关重要，我们观察每轮生成的 API 调用，并报告 API 调用的宏平均精确率（P）、召回（R）和 F1 分数（F1）。

在线评估 遵循 Xi 等人 [19]，，我们将模型部署至线上并进行 A/B 测试。我们为每个模型随机抽取 120 段对话。具备领域知识的标注员需要对每段对话进行满意度评估，将满意度分为“低”“中”或“高”三个等级。[2]

3.2 主要结果

离线评估和在线评估结果分别如表3和表2所示，由此我们得出以下观察结果：（1）大规模预训练语言模型显著提升了词元级生成模型的性能。例如，与普通长短期记忆网络模型相比，T5 在 BLEU-4 分数上绝对提升了 26.20%，在 F1 上提升了 4.98%。（2）与 CDiaGPT 和 T5 模型相比，我们的轻量级 DTA 在离线评估中实现了具有竞争力的性能，并在在线评估中获得了最高的满意度评分，验证了我们所提出方法的有效性。

模型	Low	中型	High
LSTM	30.00	29.17	40.83
Transformer	27.50	28.33	44.17
CDiaGPT	12.50	13.33	74.17
T5	5.83	15.83	78.33
DTA	8.33	12.50	79.17

表2：在线评估统计结果（%）

模型	BLEU-4	P	R	F1
LSTM	20.62	66.16	78.84	71.72
Transformer	26.21	62.59	75.90	64.74
CDiaGPT	42.54	**71.18**	86.43	77.11
T5	**46.83**	70.91	87.25	76.70
DTA	44.82	68.03	**90.80**	**77.74**

表3：离线评估的统计结果（%）。

3.3 深入分析

3.3.1 效率问题的影响。为了考察效率问题的影响，我们在相同的基础设施下（即 Tesla v100、32GB 随机存取存储器等），收集了各模型处理测试样本所需的计算时间。考虑到计算时间与解码步骤高度相关，我们首先根据回复长度将生成的回复分为 10 个子集，然后计算每个子集的平均计算时间，如图 3 所示。我们可以观察到：（1）尽管性能相当，DTA 在计算效率方面显著优于其他模型（例如，子集“[50, 59]”中，DTA 为 3.37 毫秒，而 CDiaGPT 为 1265.56 毫秒，T5 为 2470.69 毫秒）。（2）对于较长的回复，DTA 相较于其他模型的优势更加显著。原因在于，词元级生成模型的解码步骤数与回复长度相同，而 DTA 执行的是动作级解码。上述观察结果验证了我们的系统能够以有限的计算资源为线上对话服务提供有效的构建方案。

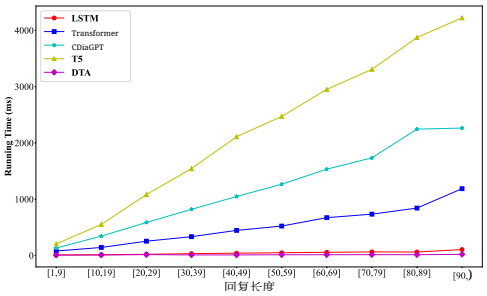


图 3：不同模型的计算时间。

3.3.2 可靠性问题的影响。我们通过定量考察重复问题，研究可靠性问题的影响。具体而言，我们计算每一轮回复与前几轮回复的 Jaccard 指数 [3]。各模型的平均 Jaccard 指数见表 4。从中可以看出：(1) 人工回复的 Jaccard 指数最低，这表明现有模型在减少重复问题方面仍有改进空间。(2) 与 CDiaGPT 和 T5 相比，我们的方法的 Jaccard 指数显著更低。行为序列生成与采样策略相结合，能够有效缓解重复问题。

模型	Jaccard 指数
人工回复	0.129
CDiaGPT	0.214
T5	0.207
DTA	0.142

表 4：不同模型的 Jaccard 指数。

图 1（b）展示了一段具体的在线对话，其中 CDiaGPT 生成的回复完全相同。尽管 DTA 生成了相似的动作序列（例如，操作 A226、A372、A249和A109的组合），但动作序列与先前轮次完全重复的情况要少得多。动作序列中的细微差异可能会导致口头响应发生较大变化。此外，采样机制确保不同轮次中的相同操作对应不同的片段（例如，三轮中的 A226），这进一步使 DTA 具有更好的多样性。

4 结论

本文提出了一种通过动作级生成构建的面向任务的对话系统。我们提出了一种有效的框架，只需最少的人工投入，即可利用大规模对话构建生成模型。实验分析表明，我们的系统能够解决现有端到端生成方法所面临的可靠性和效率问题。未来，我们有兴趣探索一种集成系统，通过端到端架构整合 DTA 中的离散模块。

[2]评分标准概括如下：(i)“低”表示聊天机器人无法正确处理用户需求。(ii)“中”表示聊天机器人能够正确处理用户需求，但可能生成不流畅或不完整的回复。(iii)“高”表示聊天机器人能够正确处理用户需求，并圆满完成对话。

5 演讲者简介

演讲者：Yuncheng Hua。他是美团的一名算法工程师，专注于研究和构建对话系统。

6 公司简介

美团是中国领先的购物平台，提供本地消费品和零售服务，包括娱乐、餐饮、配送、旅行及其他服务。

参考文献

[1] Anish Acharya、Suranjit Adhikari、Sanchit Agarwal、Vincent Auvray、Nehal Belgamwar、Arijit Biswas、Shubhra Chandra、Tagyoung Chung、Maryam Fazel- Zarandi、Raefer Gabriel、Shuyang Gao、Rahul Goel、Dilek Hakkani-Tur、Jan Jez- abek、Abhay Jha、Jiun-Yu Kao、Prakash Krishnan、Peter Ku、Anuj Goyal、Chien-Wei Lin、Qing Liu、Arindam Mandal、Angeliki Metallinou、Vishal Naik、Yi Pan、Shachi Paul、Vittorio Perera、Abhishek Sethi、Minmin Shen、Nikko Strom 和 Eddie Wang。2021年。Alexa Conversations：一种用于构建面向任务的对话系统的可扩展数据驱动方法。载于第*2021*年计算语言学协会北美分会年会：人类语言技术：系统演示论文集。计算语言学协会，线上，125 132 10 18653 1 2021 15 – 。[2] https://doi.org/ 。/v / .naacl-demos。Bill Byrne、Karthik Krishnamoorthi、Saravanan Ganesh 和 Mihir Kale。. TicketTalk：迈向端到端、基于交易的对话系统的人类水平性能。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议（第*1*卷：长篇论文）论文集。计算语言学协会，线上，671 680 10 18653 1 2021 55 – 。[3] https://doi.org/ 。/v / .acl-long。Luciano da F Costa。. Jaccard 指数的进一步推广。*arXiv* 预印本 *arXiv:2110.09619*（2021）。[4] Zihao Fu、Wai Lam、Anthony Man-Cho So 和 Bei Shi。2021年。文本生成中重复问题的理论分析。载于*AAAI* 人工智能会议论文集，第35卷。12848–12856。[5] Youngsoo Jang、Jongmin Lee 和 Kee-Eung Kim。2021年。GPT-Critic：用于端到端面向任务的对话系统的离线强化学习。载于国际学习表征会议。[6] Meixun Jin、Mi-Young Kim、Dongil Kim 和 Jong-Hyeok Lee。2004年。使用逗号分割中文长句。载于第三届 *SIGHAN* 中文处理研讨会论文集。1–8。[7] Sungjin Lee、Qi Zhu、Ryuichi Takanobu、Zheng Zhang、Yaoqin Zhang、Xiang Li、Jinchao Li、Baolin Peng、Xiujun Li、Minlie Huang 和 Jianfeng Gao。2019年。ConvLab：多领域端到端对话系统平台。载于第*57*届计算语言学协会年会：系统演示论文集。计算语言学协会，意大利佛罗伦萨，64–69。https://doi.org/ 10 18653 1 19 3011 。/v /P - Feng-Lin Li、[8] Minghui Qiu、Haiqing Chen、Xiongwei Wang、Xing Gao、Jun Huang、Juwei Ren、Zhongzhou Zhao、Weipeng Zhao、Lei Wang 等。2017年。Alime assist：一款用于打造创新电子商务体验的智能助手。载于*2017*年 *ACM* 信息与知识管理会议论文集。2495–2498。[9] Zhaojiang Lin、Andrea Madotto、Genta Indra Winata、Peng Xu、Feijun Jiang、Yuxiang Hu、Chen Shi 和 Pascale Fung。2021年。Bitod：用于面向任务的对话建模的双语多领域数据集。*arXiv* 预印本 *arXiv:2106.02787*（2021）。[10] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019年。语言模型是无监督的多任务学习器。*OpenAI* 博客1，8（2019），9。[11] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020年。探索统一文本到文本转换器的迁移学习极限。机器学习研究期刊21，140（2020），1–67。[12] Abigail See、Stephen Roller、Douwe Kiela 和 Jason Weston。2019年。怎样才算一场好的对话？可控属性如何影响人类判断。*arXiv* 预印本 *arXiv:1902.08654*（2019）。[13] Shuangyong Song、Chao Wang、Haiqing Chen 和 Huan Chen。2021年。一种用于提升电子商务客服聊天机器人用户满意度的情感安抚框架。载于第*2021*年计算语言学协会北美分会年会：人类语言技术：行业论文集。计算语言学协会，线上，130–137。https://doi.org/10.18653/v1/2021.naacl-industry.17[14] Kai Sun、Seungwhan Moon、Paul Crook、Stephen Roller、Becka Silvert、Bing Liu、Zhiguang Wang、Honglei Liu、Eunjoon Cho 和 Claire Cardie。2021年。加入闲聊以增强面向任务的对话。载于第*2021*年计算语言学协会北美分会年会论文集：人类语言技术。计算语言学协会，线上，1570 1583 10 18653 1 2021 124 – 。https://doi.org/ 。/v / .naacl-main。Ilya Sutskever、Oriol Vinyals 和 Quoc V Le。. 使用神经网络进行序列到序列学习。神经信息处理系统进展27（2014）。[16] Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kul- shreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du 等。2022年。Lamda：对话应用语言模型。*arXiv* 预印本 *arXiv:2201.08239*（2022）。[17] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是你所需要的一切。神经信息处理系统进展30（2017）。[18] Yida Wang、Pei Ke、Yinhe Zheng、Kaili Huang、Yong Jiang、Xiaoyan Zhu 和 Minlie Huang。2020年。大规模中文短文本对话数据集。载于*NLPCC*。https://arxiv.org/abs/2008.03946[19] Xiangyu Xi、Chenxu Lv、Yuncheng Hua、Wei Ye、Chaobo Sun、Shuaipeng Liu、Fan Yang 和 Guanglu Wan。2022年。一款低成本、可控且可解释的面向任务聊天机器人：以真实世界售后服务为例。载于第*45*届国际计算机协会信息检索研究与开发会议论文集（西班牙马德里）*(SIGIR '22)*。计算机协会，纽约，美国，3398–3402。https://doi.org/10.1145/3477495. 3536331[20] Yuanmeng Yan、Rumei Li、Sirui Wang、Fuzheng Zhang、Wei Wu 和 Weiran Xu。2021年。ConSERT：用于自监督句子表征迁移的对比学习框架。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议（第*1*卷：长篇论文）论文集。计算语言学协会，线上，5065 5075 10 18653 1 2021 393 – 。[21] https://doi.org/ 。/v / .acl-long。Zhao Yan、Nan Duan、Peng Chen、Ming Zhou、Jianshe Zhou 和 Zhoujun Li。2017年。构建面向在线购物的任务型对话系统。载于第三十一届 *AAAI* 人工智能会议。[22] Yunyi Yang、Yunhao Li 和 Xiaojun Quan。2021年。Ubar：借助 GPT-2 实现完全端到端的面向任务对话系统。载于*AAAI* 人工智能会议论文集，第35卷。14230–14238。[23] Dian Yu、Luheng He、Yuan Zhang、Xinya Du、Panupong Pasupat 和 Qi Li。2021年。利用检索示例进行少样本意图分类和槽位填充。载于第*2021*年计算语言学协会北美分会年会：人类语言技术论文集。计算语言学协会，线上，734–749。https://doi.org/10.18653/v1/2021. naacl-main.59[24] Qi Zhu、Zheng Zhang、Yan Fang、Xiang Li、Ryuichi Takanobu、Jinchao Li、Baolin Peng、Jianfeng Gao、Xiaoyan Zhu 和 Minlie Huang。2020年。ConvLab-2：用于构建、评估和诊断对话系统的开源工具包。载于第*58*届计算语言学协会年会：系统演示论文集。计算语言学协会，线上，142 149 10 18653 1 2020 19 – 。https://doi.org/ 。/v /.acl-demos。Xiaoming Zhu。. 案例 ii（部分 a）：Jimi 的成长之路：人工智能重新定义了京东的客户服务。载于数字经济中的新兴冠军。Springer，91–103。[25] 2014 2019