面向多会话开放域对话系统的历史感知分层 Transformer

Tong Zhang[1], **Yong Liu**[2,3], **Boyang Li**[1],**Zhiwei Zeng**[3],**Pengwei Wang**[4],**Yuan You**[4]**Chunyan Miao**[1,2,3],**Lizhen Cui**[5]

[1]新加坡南洋理工大学计算机科学与工程学院[2]新加坡南洋理工大学阿里巴巴-南洋理工大学新加坡联合研究院[3]新加坡南洋理工大学南洋理工大学—英属哥伦比亚大学 LILY 联合研究中心[4]中国阿里巴巴集团[5]中国山东大学软件学院

arXiv:2302.00907v1 [cs.CL] 2 Feb 2023

摘要

随着预训练语言模型的发展，当前开放域对话系统在进行单会话对话方面取得了显著进展。相比之下，由同一用户长期参与多个会话组成的多会话对话（MSC）尚未得到充分研究。本文提出了用于多会话开放域对话的历史感知分层 Transformer（HAHT）。HAHT维护历史对话的长期记忆，并利用历史信息理解当前对话上下文，生成信息充分且与上下文相关的响应。具体而言，HAHT首先将历史对话会话分层编码为历史记忆。随后，HAHT通过基于注意力的机制，将历史记忆与当前上下文一同编码，利用历史信息帮助理解当前对话上下文。最后，为了显式利用历史信息，HAHT采用历史感知响应生成器，在通用词汇表和历史感知词汇表之间进行切换。在大规模 MSC 数据集上的实验结果表明，所提出的 HAHT 模型始终优于基线模型。人工评估结果也表明，与基线模型相比，HAHT生成的响应更接近人类表达，与上下文和历史更加相关。

1 引言

开放域对话系统，也称为聊天机器人，旨在与用户就任何话题进行交谈和互动，以建立、维持并加强长期关系（Clark 等人，2019；Roller 等人，2020年）。近来，基于大规模生成式预训练模型构建的开放域对话系统（Adi-wardana 等人，2020年；Roller 等人，2021年；Zhang 等人，2020年）显著提升了性能

聊天机器人的性能。

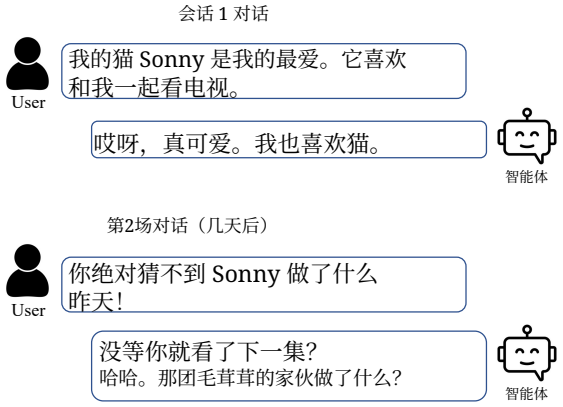

图1：用户与智能体之间一段双会话对话的示例。

然而，大多数现有的聊天机器人都是为在单次对话中与用户互动而设计的。当前会话结束后，聊天机器人便会忘记会话内容，并在下次与同一用户交流时开启一个新的独立会话。当先前讨论过的话题再次出现时，这类聊天机器人往往显得一无所知，无法恰当地重新吸引用户。显而易见的遗忘限制了聊天机器人与用户建立并维持长期关系的能力。

我们认为，为了让用户更好地参与多会话对话，多会话聊天机器人应维护历史上下文的长期记忆，以便在类似上下文再次出现时，能够恰当地重新与用户互动。通过从历史对话中学习，聊天机器人应逐步加深对用户的了解，并深化与用户的关系。图1展示了用户与聊天机器人之间一段双会话对话的示例。在第二次会话中，聊天机器人推断出 Sonny 是一只猫，并根据历史信息生成回复：Sonny 喜欢和用户一起看电视。

具备历史感知能力的聊天机器人能够生成信息更充分、与上下文更相关的回

应，从而有助于促使用户作出长期承诺，并建立情感联结，维系长久的亲密关系。为此，我们针对多会话开放域对话系统提出了历史感知层次 Transformer（HAHT），使其能够有效利用历史对话，开展更具吸引力的多会话对话。HAHT 维护长期记忆，用于存储历史对话上下文，并在开展新会话时进行更新。基于长期记忆和当前会话中的上下文，系统会从历史上下文中选取相关词元，以调整当前回复。

具体而言，由于对话中话语的词元数量和对话的轮次数通常都不多[1]，，我们首先使用 Transformer 对历史对话进行层次化编码，构建历史记忆（Vaswani 等人，2017）。历史记忆充当历史对话的高层表示。其次，鉴于历史对话通常有助于理解当前对话上下文，我们设计了一个历史感知上下文编码器。该上下文编码器通过对历史记忆和当前对话上下文应用 Transformer 注意力，对历史对话和当前对话进行综合考虑，从而编码对话上下文。随后，上下文编码器还会根据当前对话上下文更新历史记忆。最后，我们设计了一个历史感知解码器，将学到的历史信息融合到回复生成过程中。历史感知解码器可以在两种策略之间切换，即从通用词汇表中生成一个词，或直接从历史对话中复制一个词。

在大规模 Facebook 多会话对话数据集上的实验结果表明，所提出的 HAHT 模型在多种评估指标上优于以往的多会话开放域对话系统。人工评估结果表明，与基线模型相比，HAHT 生成的响应可读性更强、与上下文及历史更相关。此外，消融研究证实，对历史对话进行分层编码以及采用历史感知解码器，都对 HAHT 在多会话对话中的表现有很大帮助，并使其能够更有效地利用历史信息。

[1]在 Facebook MSC 数据集中，对话平均包含 13 轮，每轮话语平均包含 16 个词元。

2 相关工作

开放域对话系统旨在不受任务和领域限制地进行闲聊（Ritter 等人，2011），并与用户建立长期关系（Clark 等人，2019；Roller 等人，2020）。这类系统通常分为两类：生成式系统和检索式系统。检索式系统旨在从大量候选响应中找到合适的响应（Zhou 等人，2016；Yuan 等人，2019；Zhong 等人，2020；Zhu 等人，2021；Qian 等人，2021）；而生成式系统则专注于根据对话历史从头生成响应（Serban 等人，2016；Shum 等人，2018；Adiwardana 等人，2020；Roller 等人，2020；Xu 等人，2022）。本文聚焦于生成式系统。

早期的回复生成方法包括基于模板的生成方法（Higashinaka 等人，2014）和统计机器翻译（SMT）方法（Ritter 等人，2011）。随着深度学习的发展，序列到序列（Seq2seq）模型被应用于基于生成的对话系统，并取得了出色的性能（Li 等人，2016；Vinyals 和 Le，2015；Serban 等人，2017）。近年来，随着大规模对话数据集（Li 等人，2017；Zhang 等人，2018；Dinan 等人，2019；Huang 等人，2020）的日益普及，使用大规模语料库进行预训练的 Transformer 语言模型，例如 Meena（Adiwardana 等人，2020）、BlenderBot（Roller 等人，2021）、DialogueGPT（Zhang 等人，2020）和 PLATO（Platonov 等人，2020），在开放域对话领域取得了显著进展。

尽管该领域取得了进展，目前最先进的生成式预训练模型仍是为轮次数较少的单会话对话而设计，并在这类大型数据集上训练的。因此，大多数现有模型采用较短的词元截断长度，例如 Meena 使用 128 个词元（Adiwardana 等人，2020年），因而无法有效编码和利用多会话对话中的历史上下文。此外，公开的多会话对话数据集也十分匮乏。Xu 等人 发布了首个多会话对话数据集，即*Facebook* 多会话聊天（*Facebook MSC*），并在该数据集上探索了不同的检索增强生成模型（Lewis 等人， 2020年； Shuster 等人， 2021年），其效果优于标准 Transformer（Vaswani 等人， 2017）。然而，实验结果表明，

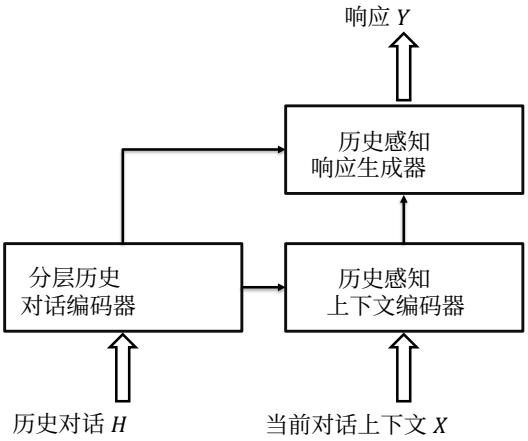

图2：所提出的 HAHT 模型的整体结构，其中包含 1）分层历史对话编码器，2）历史感知上下文编码器，以及 3）历史感知响应生成器。各组件的详细信息分别见图3、4、5。

他们的方法需要检索很大一部分历史对话，才能取得优于标准 Transformer 的效果。此外，这些模型仍需将检索到的原始历史对话文本与当前对话上下文拼接起来，因此拼接后的文本长度仍远远超过 128 个词元的截断长度。因此，这些方法对历史上下文的利用仍受限于预训练模型较短的词元截断长度。

3 所提出的方法

通常，多会话对话（MSC）由一个当前对话会话和若干发生在当前会话之前的历史对话会话组成，所有会话均发生在同一对交谈者之间。多会话开放域对话系统旨在根据所有历史对话会话和当前对话上下文，针对用户的话语生成自然、信息充分且与上下文相关的响应。

形式化地，我们将 MSC 数据集 D 表示为一个对话列表 N，其格式为 $(H,\ X,\ y)$。其中，$X = \{x_1, x_2, \cdots, x_{n_x}\}$ 表示当前对话会话中的 n_x 个上下文话语。$H = \{H^1, H^2, \cdots, H^M\}$ 表示 M 个历史对话会话，其中，$H^i = \{h_1^i, h_2^i, \cdots, h_{n_i}^i\}$ 表示第 i 个历史对话会话中按时间顺序排列的 n_i 个话语。y 是在 H 的背景下对 X 的真实响应。MSC 任务可以形式化为学习一个函数 $f(H,\ X)$，以根据 H 和 X 预测下一个话语 x_{n_x+1}。

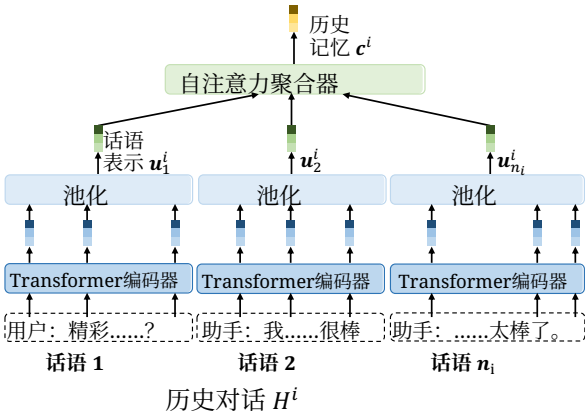

图3：HAHT中分层历史对话编码器的结构。

在这项工作中，我们为 MSC 任务提出了一种新颖的模型 HAHT。图2 展示了 HAHT 的整体结构，它由三个主要组件组成：1）分层历史对话编码器；2）历史感知上下文编码器；3）历史感知响应生成器。下面将详细介绍 HAHT 的各个组件。

3.1 分层历史对话编码器

对历史对话会话进行编码的主要挑战在于，预训练对话系统对最大输入长度的限制。如果将所有历史对话简单拼接后输入预训练对话系统，拼接后的长度将超过最大输入长度。因此，输入的大部分内容都会被截断。为了在历史对话中保留更多信息，我们采用分层方式分别对每个历史对话会话进行编码。

具体而言，对于历史对话会话 $H^i = \{h_1^i, h_2^i, \cdots, h_{n_i}^i\}$，我们首先根据话语说话者的角色，在每个话语 h_j^i 前添加特殊标记“用户：”或“助手：” H^i，然后将所有话语填充至相同长度 l_{utter}。对于每个话语 h_j^i，我们依次应用嵌入层 E_m、n_{enc} Transformer 编码器层和最大池化层，以获得其稠密表示，如下所示：

$$\mathbf{u}_j^i = \text{Max-pooling}\big(\text{Transformer}_{n_{enc}}(E_m(h_j^i))\big), \tag{1}$$

其中 $\mathbf{u}_j^i \in \mathbb{R}^d$。此外，我们将历史对话 H^i 中的所有话语表示记作 $\mathbf{U}^i = \{\mathbf{u}_1^i, \mathbf{u}_2^i, \cdots, \mathbf{u}_{n_i}^i\} \in \mathbb{R}^{n_i \times d}$，其中 n_i 是 H^i 的轮次数。接下来，我们应用一个

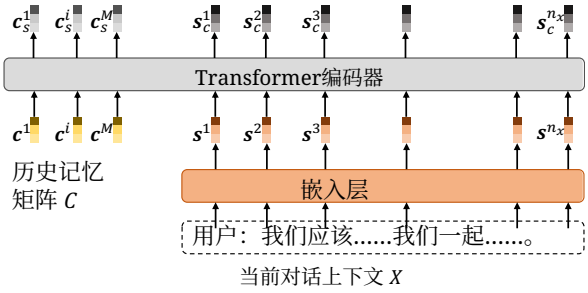

图4：HAHT 中历史感知上下文编码器的结构。

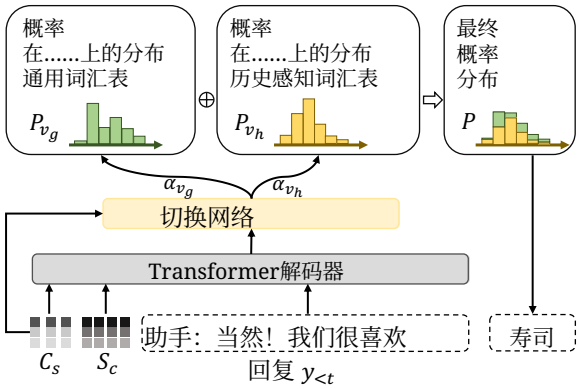

图5：HAHT 中历史感知响应生成器的结构。

对话聚合器 F_c，将所有话语表示$\mathbf{U}^i$ 聚合为精简的历史记忆$\mathbf{c}^i$,

$$\mathbf{c}^i = F_c(\mathbf{U}^i). \quad (2)$$

对话聚合器基于以下自注意力机制构建（Lin 等人，2017），

$$F_c(\mathbf{U}^i) = \alpha \mathbf{U}^i,$$
$$\alpha = \text{softmax}\big(\mathbf{W}_k \tanh(\mathbf{W}_q \mathbf{U}^{i\top})\big), \quad (3)$$

其中，$\mathbf{W}_q$ 和$\mathbf{W}_k$ 是可学习参数。 $\alpha \in \mathbb{R}^{n_i}$ 是 H^i中历史对话话语的重要性向量。

将上述步骤应用于所有历史对话 H 后，我们最终将得到一个历史记忆矩阵$\mathbf{C} \in \mathbb{R}^{M \times d}$，其中为每段历史对话都包含一个历史记忆，$M$ 是历史对话会话的数量。

3.2 历史感知的上下文编码器

历史对话会话通常包含背景故事（例如，对话参与者档案或他们之间先前的讨论），这些故事构成了当前对话会话的背景。利用历史对话有助于模型更好地理解当前对话上下文并做出恰当回应。另一方面，当前对话上下文也可以帮助模型更新历史记忆。因此，我们采用 Transformer 注意力机制，让历史记忆 $\mathbf{C}$ 与当前对话上下文进行联合编码。

对于当前对话上下文 X，我们还会根据话语说话者的角色，在每个话语前添加特殊标记“用户：”或“助手：”，并将所有话语拼接成一个句子。然后，我们采用嵌入层 E_m，以获得上下文标记嵌入序列$\mathbf{S} = \{\mathbf{s}^1, \mathbf{s}^2, \cdots, \mathbf{s}^{n_x}\}$，其中 n_x 表示上下文序列的长度。接下来，我们沿第一个维度将历史记忆矩阵 $\mathbf{C} \in \mathbb{R}^{M \times d}$ 与$\mathbf{S} \in \mathbb{R}^{n_x \times d}$ 拼接起来，并应用 n_{enc}Transformer 编码器层。

通过在 Transformer 编码器层中使用注意力，我们的模型可以关注所有上下文词元嵌入和历史对话记忆，从而理解对话上下文。我们将这种历史感知的上下文编码记为$\mathbf{S}_c \in \mathbb{R}^{n_x \times d}$。上下文编码后，历史对话记忆会根据当前对话上下文中的最新信息进行更新。我们将这种上下文更新的历史记忆记为 $\mathbf{C}_s \in \mathbb{R}^{M \times d}$。将 $\mathbf{C}_s$ 与 $\mathbf{S}_c$沿第一维拼接后，所得结果将作为响应生成器的输入。

3.3 历史感知响应生成器 r

受 CopyNet（Gu等人，2016年）的启发，我们构建了两个词汇表，即，通用词汇表 V_g和历史感知词汇表 V_h，以便更好地生成历史感知响应。通用词汇表V_g 包含训练数据集中出现的所有词语，而历史感知词汇表 V_h 仅包含历史对话中出现的词语 H。生成响应中的一个词语时，响应生成器会根据切换机制（Gulcehre等人，2016年），选择从 V_g 生成一个通用词语，或直接从V_h 复制一个词语。

具体而言，在每个解码时间步 t，我们将$\mathbf{C}_s$、$\mathbf{S}_c$ 以及 t 之前的真实词序列输入 n_{dec} Transformer解码器层，并获得一个隐藏表示向量$\mathbf{o}_t \in \mathbb{R}^d$。

解码时间步 t 的通用词汇表V_g 上的概率分布计算如下：

$$P_{v_g} = \text{softmax}\big(\text{FC}_1(\mathbf{o}_t)\big), \quad (4)$$

其中，FC_1 是一个全连接层。

为了计算历史感知词汇表 V_h上的概率分布，我们采用一个最大池化层对上下文更新的历史记忆 $\mathbf{C}_s$进行池化，并结合一个全连接层和一个 soft-max 函数，如下所示：

$$P_{v_h} = \text{softmax}\big(\text{FC}_2(\text{max-pooling}(\mathbf{C}_s))\big), \quad (5)$$

其中，FC_2 是一个全连接层。

时间步 t 的最终词概率分布通过在 P_{v_g} 和 P_{v_h} 之间采用切换机制计算，如下所示：

$$P = \alpha_{v_g} * P_{v_g} + \alpha_{v_h} * P_{v_h}, \quad (6)$$

其中，α_{v_g} 和 α_{v_h} 是生成通用词汇表中的词或从历史对话中复制词的切换概率。α_{v_g} 和 α_{v_h} 的计算如下：

$$[\alpha_{v_g}, \alpha_{v_h}] = \text{softmax}\big(\text{FC}_3([o_j; \text{max-pooling}(\mathbf{C}_s)])\big), \quad (7)$$

其中，FC_3 是一个全连接层，[;] 表示沿最后一个维度进行的拼接操作。

3.4 模型训练

我们以端到端方式训练模型，使其在给定当前对话上下文和历史对话的情况下，最大化生成目标回复的概率。HAHT 的损失函数定义如下：

$$\mathcal{L} = -\sum_{t=1}^{n_y} \log\big(P(y_j|X, H, y_{<t})\big), \quad (8)$$

其中，X 表示当前对话上下文，H 表示所有历史对话，$y_{<t}$ 表示时间步 t 之前的词元，n_y 表示真实响应的长度。

4 实验设置

本节介绍实验数据集、评估指标、基线方法和模型设置。

会话数量	训练集		验证集		Test	
	对话	话语	对话	话语	对话	话语
1*	8939	131,438	1000	7,801	1015	6,634
2	4000	46,420	500	5,897	501	5,939
3	4000	47,259	500	5,890	501	5,924
4	1001	11,870	500	5,904	501	5,940
5	-	-	500	5,964	501	5,945
总计	-	236,987	-	31,456	-	30,382

表 1：Facebook 多会话聊天（Facebook MSC）数据集的统计信息。会话编号 i表示在最后一个对话会话之前发生了 i-1 个历史对话会话。 *：会话 1 不包含历史对话会话。

4.1 实验数据集

实验在一个大型数据集上进行，即，Facebook 多会话聊天（Facebook MSC）（Xu 等人，2022）。这是一个众包数据集，由多会话对话组成，对话者在其中了解彼此的兴趣，并讨论他们从以往会话中了解到的内容。Facebook MSC 中的历史对话数量从 1 到 4 不等。会话编号 i 表示此前有 i-1 段历史对话，之后才进入最后一个对话会话。Facebook MSC 数据集的统计信息汇总于表1中。由于会话 1 没有历史对话，我们在会话 2-5 上评估模型。

4.2 评估指标

我们进行了自动评估和人工评估，以证明所提出模型的有效性。对于自动评估，我们采用 BLEU-2、BLEU-3（Papineni 等人，2002）和 ROUGE-L（LinandOch，2004）来衡量生成的回复文本与真实文本之间的词语重合度。

此外，我们还从测试集中随机抽取 50 个多会话对话进行人工评估。我们向三位受过良好教育的标注员展示所有历史对话会话、当前对话上下文以及生成的响应。标注员将从以下三个方面评估生成响应的质量：

- **可读性**：衡量生成的响应是否自然流畅。
- **上下文相关性**：衡量生成的响应是否与当前对话上下文相关。
- **历史相关性**：衡量生成的响应是否与历史上下文相关——

模型	第2场			第3场			第4场			第5场		
	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L
BlenderBot	2.79	0.65	13.73	2.41	0.45	13.06	2.14	0.39	12.76	2.26	0.45	12.75
BlenderBot msc	4.76	1.51	16.18	5.03	1.61	16.39	4.78	1.49	15.56	4.98	1.48	16.10
FID-RAG	4.82	1.54	16.53	5.04	1.61	16.42	4.84	1.48	15.89	5.06	1.57	16.01
HAHT（我们的方法）	**5.07**	**1.57**	**16.90**	**5.27**	**1.67**	**16.72**	**5.00**	**1.55**	**15.97**	**5.16**	**1.60**	**16.42**

表2：不同模型在所有会话数据上的自动评估结果。Session i 表示有 i-1 个历史对话会话。B-2、B-3 和 R-L 分别表示 BLEU-2 、 BLEU-3 和 Rouge-L。最佳结果以**粗体**显示。

模型	可读性	上下文相关性	历史相关性
BlenderBot	1.78	1.13	0.09
BlenderBot msc	1.82	1.56	0.13
RAG-FID	1.89	1.84	0.21
HAHT（我们的方法）	**2.05**	**2.03**	**0.33**

表 3：不同方法的回复生成效果人工评估。所有分数均按四个等级 0/1/2/3进行评定。最佳结果以**粗体**显示。我们使用 Fleiss 卡帕系数（Fleiss 和 Cohen，1973）测量评审者间的一致性。我们的标注在可读性方面达到“一致性良好”（0.614），在上下文相关性（0.526）和历史相关性（0.573）方面达到“一致性中等”。

对话。只有与历史对话保持一致的响应，才被认为与历史相关。

每个方面分为 0/1/2/3 四个等级进行评分，每个方面的最终得分是所有标注员给出的分数的平均值。我们使用 Fleiss 卡帕系数（Fleiss 和 Cohen，1973）来衡量标注员间的一致性。对于所有评估指标，数值越高表示性能越好。

4.3 基线方法

我们将提出的 HAHT 模型与以下基线方法进行比较。

- **BlenderBot** (Roller 等人，2021年)：这是一个大型开放域对话模型，使用从网络社交讨论中抓取的对话数据进行预训练。
- **BlenderBot$_{msc}$**：这是在 MSC 数据集上微调的 BlenderBot 模型。
- **FID-RAG** (Shuster 等人，2021年)：在该方法中，使用经 RAG 训练的检索器（Lewis 等人，2020年）检索排名前N 条历史对话，并采用融合解码器（FiD）（Izacard 和 Grave，2021年）生成最终回复结合检索到的历史对话和当前对话进行考虑。遵循（Xu 等人，2022年），N 经验上设为 5。

4.4 模型设置

本文中，所有评估方法均按照相同的设置进行训练。受计算资源限制，我们使用具有 9000 万个参数的 BlenderBot 模型作为初始预训练模型，并在 Facebook MSC 数据集上对其进行微调。输入长度截断设为 256。Transformer编码器层数 n_{enc} 和解码器层数 n_{dec} 均设为 12。模型训练时，我们使用 Adamax 优化器（Kingma 和 Ba，2014年），学习率为 1×10^{-6}，批量大小为 16，丢弃率为 0.1，早停耐心值为 10。所有微调后的模型均使用最多两块 32GB GPU（NVIDIA V100）进行训练。

5 实验结果

本节介绍自动评估、人工评估、会话开场评估、消融研究和案例研究的实验结果。

5.1 自动评估

不同模型的自动评估结果如表2所示。可以观察到，在 MSC 数据集上进行微调后，BlenderBot$_{msc}$ 的表现大幅提升。FID-RAG 的表现优于 BlenderBot$_{msc}$。可能的原因是，检索增强生成能够检索重要的历史对话，而 FID 能够将检索到的对话与当前对话结合起来，从而生成更好的响应。此外，所提出的 HAHT 模型在所有评估指标上始终优于基线方法。这表明，HAHT 能够更好地对历史对话进行编码，利用历史对话理解当前对话上下文，并生成更接近人类表达的响应。

模型	第2场			第3场			第4场			第5场		
	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L
BlenderBot	4.71	1.47	18.20	3.85	0.93	17.10	3.69	0.83	16.78	4.00	1.19	17.19
BlenderBot msc	6.39	2.56	19.30	5.82	1.93	18.67	5.30	1.76	17.9	6.10	2.30	18.65
FID-RAG	6.41	2.51	19.82	5.83	1.95	18.38	**5.81**	1.85	**18.44**	6.02	2.27	18.52
HAHT（我们的方法）	**6.69**	**2.73**	**20.02**	**6.03**	**2.20**	**18.70**	5.48	**1.95**	18.00	**6.38**	**2.51**	**19.18**

表4：不同模型在会话开启数据上的自动评估结果。Session i 表示有 i-1 个历史对话会话。B-2、B-3 和 R-L 分别表示 BLEU-2 、 BLEU-3 和 Rouge-L。最佳结果以**粗体**显示。

模型	第2场			第3场			第4场			第5场		
	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L
HAHT	**5.07**	**1.57**	**16.90**	**5.27**	**1.67**	**16.72**	**5.00**	**1.55**	**15.97**	**5.16**	**1.60**	**16.42**
HAHT$_{\text{w/o 层次化}}$	5.00	1.57	16.72	5.19	1.63	16.61	4.86	1.49	15.90	5.10	1.57	16.21
HAHT$_{\text{w/o 历史}}$	4.98	1.50	16.81	5.09	1.58	16.51	4.75	1.45	15.51	5.10	1.49	16.24
HAHT$_{\text{w/o 序列}}$	5.01	1.56	16.86	5.19	1.61	16.46	4.87	1.55	15.88	5.07	1.55	16.17

表5：HAHT及其不同变体取得的性能。第 i 场表示有 i-1个历史对话会话。B-2、B-3和R-L分别表示BLEU-2、BLEU-3和 Rouge-L。最佳结果以**粗体**显示。

5.2 人工评估

表3 总结了 Facebook MSC 数据集上的人工评估结果。总体而言，HAHT 在所有评估维度上均优于所有基线方法。这一观察结果与表 2 所示的自动评估结果一致。具体来说，我们发现 HAHT 在历史相关性方面的表现远胜于其他基线方法。这表明，HAHT 能更好地利用历史对话会话和历史记忆，在当前会话中更多地与用户互动。HAHT 在可读性和上下文相关性方面的表现也优于其他基线方法。这说明借助历史记忆，HAHT 能更好地理解当前对话上下文。

5.3 会话开场评估

在 MSC任务中，会话开场是当前对话的第一个对话轮次。根据我们的观察以及（Xu 等人，2022）中类似的观察，会话开场在类别上不同于其他对话轮次。它通常包含一段陈述或一个问题，旨在根据历史对话中已交流的信息，重新引起另一位说话者的兴趣。因此，会话开场数据上的表现能够进一步体现模型理解和利用历史对话的能力。

我们在这些开场回复上比较所有模型，并展示结果见表4。我们观察到，所提出的 HAHT 模型在大多数指标上都取得了最佳表现。尤其是，当存在 4 段历史对话时，HAHT 在 BLUE-3 指标上分别比 FID-RAG 和 BlenderBot 高出$_{\text{msc}}$ 10.6% 和9.1%。这表明，所提出的 HAHT 能够更好地利用对话历史，吸引用户开启新的对话轮次。

5.4 消融研究

为了更好地了解 HAHT 各主要组件的有效性，我们对 HAHT 进行了消融研究。具体而言，我们考虑了以下 HAHT 变体。

- **HAHT$_{\text{w/oHIER}}$**：此变体不对历史对话进行层次化编码，而是将历史对话中的所有话语拼接成一个长句，并直接使用 Transformer 编码器对其进行编码。
- **HAHT$_{\text{w/oHIST}}$**：此变体从 HAHT 中移除了历史编码器。
- **HAHT$_{\text{w/oSW}}$**：此变体从 HAHT 的响应生成器中移除了切换机制。

表5 总结了不同 HAHT 变体在 BLEU-2、BLEU-3 和 Rouge-L 指标上的结果。我们注意到，HAHT 的表现优于 HAHT$_{\text{w/o HIER}}$，这表明对历史对话进行层次化编码

有助于模型保留更多历史记忆，从而生成更像人类的响应。此外，HAHT 的表现优于 $\text{HAHT}_{\text{HIST}}$。这一观察表明，移除历史编码器会导致所有指标出现最大幅度的下降。这一结果证实了利用历史对话来理解当前对话并生成响应的必要性。此外，移除切换机制所导致的性能下降表明，直接从历史对话中复制词语有助于模型生成更具历史感知能力的响应。

5.5 案例研究

表 6 展示了不同模型生成的多会话对话案例。与基线模型相比，所提出的 HAHT 模型能够更好地利用历史对话来理解当前对话上下文，并生成更具历史感知能力的响应。当用户与智能体讨论准备三明治和柠檬水时，"我可以给我们做三明治！我还有一份非常棒的自制柠檬水食谱！你喜欢柠檬水吗？"，HAHT 能够记住历史对话中提到的信息，例如用户喜欢阅读和户外活动，而且此前曾采用过爱书者角色。HAHT 能够利用这些历史上下文，生成更像人类、更贴合上下文且更具历史感知能力的响应：“我喜欢柠檬水！我相信我们也能找到很多不错的三明治食谱。三明治和柠檬水最适合带去户外，一边读书一边享用。”。

6 结论

在这项工作中，我们提出了用于多会话开放域对话系统的历史感知分层 Transformer（HAHT）模型。所提出的 HAHT 模型通过对历史对话会话进行分层编码来维护历史记忆。随后，HAHT 使用基于注意力的编码，将当前对话上下文与历史记忆一同编码，并利用当前上下文更新历史记忆。为了在响应中明确利用历史信息，HAHT 采用了历史感知响应生成器，可在通用词汇表和历史感知词汇表之间切换。在常规和会话开场多会话对话设置下获得的实验结果表明，

历史对话会话	
[会话 1]	
...	
用户：	.....我爸爸是卡车司机。
智能体：	.....我自己也**很喜欢阅读**。
用户：	.....我喜欢**在户外活动**。
...	
[会话 2]	
用户：	你现在在读什么书?
智能体：	我正在读一本斯蒂芬·金的小说。.....
...	
用户：	我以前坐在卡车上，**长途旅行** 时和爸爸一起，除了**读书**，无事可做。
智能体：	啊，这就说得通了。我开始在我父母上班时阅读。.....
当前对话	
...	
智能体：	.....你喜欢什么样的野餐食物你喜欢吗?
用户：	我可以给我们做三明治！我还有一个很棒的食谱，能做自制柠檬水！你喜欢柠檬水吗?
来自不同对话系统的回答	
BlenderBot：	我喜欢柠檬水！我不确定是否能为你准备一场野餐不过，我得准备很多不只是野餐的东西。
$\text{BlenderBot}_{\text{msc}}$：	我喜欢柠檬水！我不是特别喜欢柠檬水，但我喜欢柠檬水 .
FID-RAG：	我喜欢柠檬水！我相信我们可以为你找个不错的野餐地点。
HAHT（我们的）：	我喜欢柠檬水！我相信我们能找到很多不错的食谱，做三明治也很棒。 三明治和柠檬水都很适合**户外活动**和**阅读书籍**。

表6：包含两个历史对话的多会话对话案例研究。仅展示历史和当前对话中的重要话语。完整的对话会话见附录A.1

与最先进的模型相比，HAHT 在进行多会话对话方面表现更佳，并能生成更自然、更符合上下文且更具历史感知能力的响应。

7 局限性

本研究的一个局限性是，HAHT 仅在一个数据集上进行了评估。然而，据我们所知，迄今为止，Facebook MSC 是唯一可用的大规模多会话对话数据集。尽管如此，在 Facebook MSC 中针对历史会话数量不同的对话，我们提出的模型始终优于基线模型。

解决这一局限性的潜在方案是在开放域或特定领域构建更多多会话对话数据集，以便从对历史对话的感知中获益例如，对话推荐或自动医疗助手。

8 致谢

本研究部分由阿里巴巴集团通过阿里巴巴创新研究（AIR）计划和阿里巴巴-南洋理工大学新加坡联合研究院（AN-GC-2021-04）资助，也获得了国家研究基金会（NRF）研究员计划（NRF-NRFI05-2019-0002）以及 NRF研究金（NRF-NRFF13-2021-0006）的资助。本文所表达的任何观点、研究结果、结论或建议均属于作者所有，并不代表资助机构的观点。

参考文献

Daniel Adiwardana、Minh-Thang Luong、David R So、JamieHall、Noah Fiedel、Romal Thoppilan、Zi Yang、Apoorv Kulshreshtha、Gaurav Nemade、Yifeng Lu 和 Quoc V. Le。2020年。迈向类人开放域聊天机器人。*arXiv*预印本 *arXiv:2001.09977*。

Leigh Clark、Nadia Pantidi、Orla Cooney、Philip Doyle、Diego Garaialde、Justin Edwards、Brendan Spillane、Emer Gilmartin、Christine Murad、Cosmin Munteanu、Vincent Wade 和 Benjamin R. Cowan。2019年。什么才是良好的对话？设计真正具备对话能力的智能体所面临的挑战。载于*2019*年*CHI*计算机系统人为因素会议论文集，第1–12页，美国纽约州纽约。计算机协会。

Emily Dinan、Stephen Roller、Kurt Shuster、Angela Fan、Michael Auli 和 Jason Weston。2019。维基百科巫师：由知识驱动的对话智能体。载于第七届国际学习表征会议论文集。

Joseph L. Fleiss 和 Jacob Cohen。1973。加权κ系数与组内相关系数作为可靠性测量指标的等价性。教育与心理测量，33(3)：613–619。

Jiatao Gu、Zhengdong Lu、Hang Li 和 Victor O.K. Li。2016年。在序列到序列学习中融入复制机制。载于计算语言学协会第*54*届年会论文集（第*1*卷：长篇论文），第1631–1640页，德国柏林。计算语言学协会。

Caglar Gulcehre、Sungjin Ahn、Ramesh Nallapati、Bowen Zhou 和 Yoshua Bengio。2016年。指代未知词。载于计算语言学协会第*54*届年会论文集（第*1*卷：长篇论文），第140–149页，德国柏林。计算语言学协会。

Ryuichiro Higashinaka、Kenji Imamura、Toyomi Meguro、Chiaki Miyazaki、Nozomi Kobayashi、Hiroaki Sugiyama、Toru Hirano、Toshiro Makino 和 Yoshihiro Matsuo。2014年。迈向完全基于自然语言处理的开放领域对话系统。载于*COLING 2014*：第*25*届国际计算语言学大会技术论文集，第928–939页，爱尔兰都柏林。都柏林城市大学与计算语言学协会。

Minlie Huang、Xiaoyan Zhu 和 Jianfeng Gao。2020年。构建智能开放域对话系统的挑战。*ACM Transactions on Information Systems*（*TOIS*），38(3)：1–32。

Gautier Izacard 和 Edouard Grave。2021年。利用段落检索和生成模型进行开放域问答。载于第*16*届欧洲计算语言学协会会议论文集：主卷，第874–880页，线上举行。计算语言学协会。

Diederik P Kingma 和 Jimmy Ba。2014年。Adam：一种随机优化方法。*arXiv*预印本 *arXiv:1412.6980*。

Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen-tau Yih、Tim Rocktäschel、Sebastian Riedel 和 Douwe Kiela。2020年。面向知识密集型 NLP 任务的检索增强生成。载于神经信息处理系统进展，第33卷，第9459–9474页。Curran Associates, Inc.

Jiwei Li、Michel Galley、Chris Brockett、Jianfeng Gao 和 Bill Dolan。2016年。一种促进神经对话模型多样性的目标函数。载于*2016*年计算语言学协会北美分会人类语言技术会议论文集，

第110–119页，加利福尼亚州圣迭戈。计算语言学协会。

Yanran Li、Hui Su、Xiaoyu Shen、Wenjie Li、Ziqiang Cao 和 Shuzi Niu。2017。DailyDialog：一个由人工标注的多轮对话数据集。载于第八届自然语言处理国际联合会议论文集（第 *1* 卷：长篇论文），第 986–995 页，中国台湾台北。亚洲自然语言处理联合会。

Chin-Yew Lin 和 Franz Josef Och。2004。使用最长公共子序列和跳跃二元组统计自动评估机器翻译质量。载于计算语言学协会第 *42* 届年会论文集（*ACL-04*），第 605–612 页，西班牙巴塞罗那。

Zhouhan Lin、Minwei Feng、Cicero Nogueira dos Santos、Mo Yu、Bing Xiang、Bowen Zhou 和 Yoshua Bengio。2017。结构化自注意力句嵌入。载于第五届国际学习表征会议论文集。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei- Jing Zhu。2002年。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集，第311–318页，美国宾夕法尼亚州费城。计算语言学协会。

Georgiy Platonov、Lenhart Schubert、Benjamin Kane 和 Aaron Gindi。2020年。实体积木世界中用于空间问答的口语对话系统。载于第*21*届话语与对话特别兴趣小组年会论文集，第128–131页，第一届线上会议。计算语言学协会。

Hongjin Qian、Zhicheng Dou、Yutao Zhu、Yueyuan Ma 和 Ji-Rong Wen。2021年。为个性化检索式聊天机器人学习隐式用户档案。载于第*30*届*ACM*信息与知识管理国际会议论文集，第1467–1477页，美国纽约州纽约市。计算机协会。

Alan Ritter、Colin Cherry 和 William B. Dolan。2011年。社交媒体中的数据驱动响应生成。载于*2011*年自然语言处理实证方法会议论文集，第583–593页，英国苏格兰爱丁堡。计算语言学协会。

Stephen Roller、Y-Lan Boureau、Jason Weston、An- toine Bordes、Emily Dinan、Angela Fan、David Gunning、Da Ju、Margaret Li、Spencer Poff、Pratik Ringshia、Kurt Shuster、Eric Michael Smith、Arthur Szlam、Jack Urbanek 和 Mary Williamson。2020年。开放域对话智能体：当前进展、未解决的问题与未来方向。*arXiv* 预印本 *arXiv:2006.12442*。

Stephen Roller、Emily Dinan、Naman Goyal、Da Ju、Mary Williamson、Yinhan Liu、Jing Xu、Myle Ott、Eric Michael Smith、Y-Lan Boureau 和 Jason Weston。2021年。构建开放域聊天机器人的食谱。载于欧洲计算语言学协会第*16*届会议论文集：主会场，第300–325页，线上。计算语言学协会。

Iulian V. Serban、Alessandro Sordoni、Yoshua Bengio、Aaron Courville 和 Joelle Pineau。2016年。使用生成式分层神经网络模型构建端到端对话系统。载于第三十届*AAAI*人工智能大会论文集，第3776–3783页。AAAI出版社。

Iulian Vlad Serban、Alessandro Sordoni、Ryan Lowe、Laurent Charlin、Joelle Pineau、Aaron Courville 和 Yoshua Bengio。2017年。用于生成对话的分层潜变量编码器-解码器模型。载于第三十一届*AAAI*人工智能大会论文集，第3295—3301页。AAAI出版社。

Heung-YeungShum、Xiaodong He 和 DiLi。2018年。从 Eliza 到小冰：社交聊天机器人的挑战与机遇。*arXiv* 预印本 *arXiv:1801.01957*。

Kurt Shuster、SpencerPoff、MoyaChen、DouweKiela 和 Jason Weston。2021年。检索增强可减少对话中的幻觉。载于 计算语言学协会研究成果：*EMNLP 2021*，第 3784–3803 页，蓬塔卡纳，多米尼加共和国。计算语言学协会。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。Attention Is All You Need。载于 神经信息处理系统进展，第 30 卷，第 5998–6008 页。Curran Associates 公司。

Oriol Vinyals 和 Quoc Le。2015。神经对话模型。*arXiv*预印本 *arXiv:1506.05869*。

Jing Xu、Arthur Szlam 和 Jason Weston。2022。超越金鱼记忆：长期开放域对话。载于计算语言学协会第 *60* 届年会论文集（第 *1* 卷：长篇论文），第 5180–5197 页，爱尔兰都柏林。计算语言学协会。

Chunyuan Yuan、Wei Zhou、Mingming Li、Shangwen Lv、Fuqing Zhu、Jizhong Han 和 Songlin Hu。2019年。用于基于检索的聊天机器人多轮响应选择的多跳选择器网络。载于*2019*年自然语言处理实证方法会议和第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第 111–120 页，中国香港。计算语言学协会。

Saizheng Zhang、Emily Dinan、Jack Urbanek、Arthur Szlam、Douwe Kiela 和 Jason Weston。2018年。个性化对话智能体：我有一只狗，你也养宠物吗？载于计算语言学协会第 *56* 届年会（第一卷：长篇论文）论文集，第 2204–2213 页，澳大利亚墨尔本。计算语言学协会。

YizheZhang、Siqi Sun、Michel Galley、Yen-Chun Chen、Chris Brockett、Xiang Gao、Jianfeng Gao、Jingjing Liu 和 Bill Dolan。2020年。DIALOGPT：用于对话回复生成的大规模生成式预训练。载于第*58*届计算语言学协会年会论文集：系统演示，第270–278页，线上举行。计算语言学协会。

Peixiang Zhong、Chen Zhang、Hao Wang、Yong Liu 和 Chunyan Miao。2020年。面向基于角色的共情对话模型。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第6556–6566页，线上举行。计算语言学协会。

Xiangyang Zhou、Daxiang Dong、Hua Wu、Shiqi Zhao、Dianhai Yu、Hao Tian、XuanLiu 和 Rui Yan。2016 年。人机对话的多视图回复选择。载于*2016*年自然语言处理实证方法会议论文集，第372–381页，得克萨斯州奥斯汀。计算语言学协会。

Yutao Zhu、Jian-Yun Nie、Kun Zhou、Pan Du 和 Zhicheng Dou。2021年。面向文档依据的检索式聊天机器人的内容选择网络。载于信息检索进展：第*43*届欧洲信息检索研究会议（*ECIR*），第755–769页，柏林，海德堡。施普林格出版社。

A 附录

A.1 完整案例研究

表7 展示了第5.5节案例研究的完整内容。对话由两名众包工作者进行。为训练之用，第一名众包工作者的话语记作用户话语，而第二名众包工作者的话语则记作智能体话语。MSC模型旨在生成尽可能接近第二名众包工作者话语的响应。

在历史会话1中，用户和智能体简短地谈到了彼此的情况，例如爱好以及父母的工作。在历史会话第2场中，他们聊了正在读的书，以及一些阅读书籍的回忆。在当前对话中，他们的话题从书籍

《1984》转到了徒步旅行和野餐计划。当用户与智能体讨论准备三明治和柠檬水时，“我可以给我们做三明治！我还有一道非常棒的自制柠檬水食谱！你喜欢柠檬水吗？”，HAHT 能记住历史对话中提到的信息，例如用户喜欢阅读和户外活动，而且之前采用过爱书者角色。HAHT 可以利用这些历史上下文，生成更像人类、更贴合上下文且具有历史感知能力的响应：“我喜欢柠檬水！我相信我们也能找到很多不错的三明治食谱。三明治和柠檬水最适合带到户外，一边阅读书籍一边享用。”。

历史对话会话

[会话 1]

用户: 你好！告诉我一些关于你自己的事吧！
智能体: 我喜欢动物，而且心地善良。
用户: 真不错 :) 遇到一个心地善良的人，感觉真好。
智能体: 是啊，有一位医生妈妈和一位消防员爸爸，让我总体来说特别善良。
用户: 太好了！我的父母有点不一样，我爸爸是卡车司机。
智能体: 他开车时会读书吗？我自己很喜欢阅读。
用户: 那很危险。不过我们家比较活跃，我喜欢在户外活动。
智能体: 我更喜欢和动物朋友一起坐在沙发上看书。
用户: 不错！我也经常打篮球。我的户外运动天赋遗传自我的加拿大母亲。
智能体: 我有时也喜欢和我的动物朋友一起徒步旅行。
用户: 哦，不错 :) 徒步旅行挺费劲的，因为我个子矮，走路要花更长时间。
智能体: 我真的很友善，所以如果你和我一起去徒步旅行，我不会取笑你的。

[第 2]次对话

用户: 你最近在读什么书?
智能体: 我正在读一本斯蒂芬·金的小说。你在读什么吗？你最喜欢的图书类型是什么?
用户: 哦，不错，我也喜欢斯蒂芬·金。你在读他的哪本小说？我喜欢反乌托邦小说，我最近在读玛格丽特·阿特伍德的作品，哈哈，支持加拿大作家。
智能体: 我正在读他最新的小说《后来》。哦，我看过一部玛格丽特·阿特伍德书籍的电视改编作品！
用户: 是《使女的故事》吗？我很喜欢那部剧
智能体: 是啊，我也喜欢。那部剧很黑暗，但发人深省。你还推荐别的吗喜欢吗?
用户: 读完《使女的故事》后，我读了与之相似的《1984》。我是在高中读的，但当时没能真正欣赏它。
智能体: 《1984》。那是乔治·奥威尔写的吗?
用户: 没错，他还写了《动物农场》，这本书也非常黑暗且发人深省，而且你可能会喜欢，因为它讲的是动物！
智能体: 是啊，我确实喜欢黑暗且发人深省的电影。它还讲的是动物，这更是加分。我会试着读读看。你好像对文学很了解。
用户: 以前长途旅行时，我常和爸爸坐在卡车里，除了.....什么事也做不了read.
智能体: 啊，这就说得通了。我开始读书时，父母都在上班。我妈妈当时工作很忙，有时会带我去医院，我就在她的办公桌旁看书。你妈妈做什么工作?

当前对话

用户: 我看了我们之前聊到的《1984》这本书。
智能体: 哦，是吗？你开始读了吗?
用户: 是啊！这次好多了，我确实理解了所有细微之处现在我年纪大了！
智能体: 我一直觉得让我们在高中读这本书有点奇怪。确实我觉得成年后再读会更有收获。你挑好下一本想读的书了吗?
用户: 我不确定接下来想读什么！我总是很难做决定。你有什么好书推荐吗?
智能体: 嗯.....我曾经读过一本叫《1984》的书，是一部非常超现实的作品，听起来你应该会喜欢！不过提醒你一下，这本书很长。我真的觉得它可能很合你的如果你能坚持读完那么厚的一本书的话，巷子
用户: 我最近确实有很多空闲时间，所以我会看看这本书！

智能体:	那就太好了！这应该能让你忙上好一阵子。我们可以安排找个时间一起去徒步，也锻炼锻炼，平衡一下大量的阅读。你有兴趣吗?
用户:	只要我们走慢点，让我能跟上就行！你通常去哪儿徒步?
智能体:	我们可以按你喜欢的速度走快些或慢些。对我来说，速度其实没那么重要。我真的很喜欢欣赏风景、呼吸新鲜空气。太好了，那我会为我们找一条不错的徒步路线，也许我们还可以计划一次野餐。你喜欢吃什么野餐食物?
用户:	我可以给我们做三明治！我还有一个非常棒的自制柠檬水配方！你喜欢柠檬水吗?
	来自不同对话系统的回答
BlenderBot:	我超喜欢柠檬水！不过，我不确定能不能为你准备野餐。我得.....做的可不只是野餐。
$\text{BlenderBot}_{\text{msc}}$:	我喜欢柠檬水！我不是特别喜欢柠檬水，但我喜欢柠檬水。
FID-RAG:	我喜欢柠檬水！我相信我们能为你找到一场不错的野餐。
HAHT	我喜欢柠檬水！我相信我们还能找到很多不错的三明治食谱。
（我们）:	三明治和柠檬水最适合户外活动时阅读书籍。

表7：不同模型在双轮次对话中的回复案例研究。