中国科学院大学

University of Chinese Academy of Sciences

硕士学位论文

人类认知过程启发的开放域对话研究

作者姓名：李泽康

指导教师：冯洋　研究员

中国科学院计算技术研究所

学位类别：工程硕士

学科专业：计算机技术

培养单位：中国科学院计算技术研究所

2022 年 6 月

Human Cognition Inspired Open-domain Dialogue

A thesis submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Master of Engineering

in Computer Technology

By

Zekang Li

Supervisor: Professor Yang Feng

Institute of Computing Technology, Chinese Academy of Sciences

June, 2022

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

开放域对话系统是指在开放的领域内进行有意义的对话，是人工智能和自然语言处理领域的一个重要的研究方向。近年来，随着大规模高质量对话数据的构建与大规模预训练语言模型的进展，开放域对话系统取得了长足的进步。在人机开放域对话交互中，由于对话过程多样性高、对话信息交换频繁、对话交互过程较长，对话开放域对话系统的表现不尽如人意。具体表现为：开放域对话系统前后文不连贯、通用回复和出现知识幻觉、前后文不一致等现象。开放域对话系统本质上希望可以像人一样进行对话交互。人类在对话认知与学习过程中，需要从对话内容中充分理解对话逻辑，从外界获取相关知识（包括视觉、听觉、文本信息），保持对话前后一致等。据此，这些现象可以归结为以下与人类对话认知过程相关的问题：对话逻辑混乱问题，对话缺乏相关背景知识问题，对话缺乏前后一致性问题。

本文根据针对这三个问题，并模拟人类在对话认知过程中的行为，展开了以下研究：

1. 引入高层语义信息流的开放域对话历史建模

在开放域对话人机交互中，由于对话过程丰富多样、对话逻辑理解较为困难，开放域对话系统经常发生逻辑混乱的现象，极大影响人机交互体验。该现象归结于开放域对话系统对对话历史的建模能力不足，不能完全理解对话历史。目前，开放域对话历史建模主要有两种方式：平铺模式与分层建模。在平铺模式中，拼接对话历史使得模型较难捕捉句子之间的关系。在分层建模中，单独编码对话语句缺失了对话上下文信息。因此，本文提出了基于高层语义信息流的开放域对话历史建模方法缓解对话逻辑混乱问题。该方法通过在低层使对话语句内充分考虑上下文信息与在高层建模对话语句间语义信息流，根据已有对话逻辑，规划未来对话内容，对对话历史进行建模。在此基础上，本文设计了 DialoFlow 对话预训练模型与流评分，在对话回复生成与对话自动评估实验上都说明本文引入高层语义信息流的对话历史建模方法的有效性。

2. 引入多模态知识的开放域对话回复生成

人们在对话过程中会从周围的环境中获取多种模态的信息，例如视觉信息、

听觉信息、文本信息。这些多模态信息是相辅相成的，帮助人们理解知识和进行对话。对话机器人如果缺少相关的知识，就会产生通用回复或错误的回复。为了使开放域对话系统具备理解多模态信息的能力，本文在开放域对话系统中引入视频、音频、文本等多模态知识，主要针对多模态对话中存在的数据量稀缺问题，模态融合难问题，提出了基于预训练语言模型的通用的多模态 Transformer 模型，设计了多种多模态融合训练任务。实验表明，本文提出的通用的多模态 Transformer 模型可以有效的在开放域对话系统中融合多模态信息，生成流畅的、有信息量的对话回复。

3. 引入问询的开放域对话系统前后文一致性自动评估

目前，开放域对话系统在人机交互过程中可以生成出流畅的、多样的、有信息量的、与对话历史相关的回复，但是在对话前后一致性上还存在很多不足之处，尤其在讨论关于观点或事实时经常产生前后文冲突的现象。在人类对话交互过程中，前后文不一致会很大程度影响对话的体验，前后一致性是开放域聊天机器人的基本要求之一。提升开放域聊天机器人前后一致性需要一种有效、高效的评估方式，而当前对话前后一致性评估极度依赖耗时耗力不可复现的人工评测，并没有一种有效、高效、稳定的自动评估方法。因此，本文模拟人类在对话中反复询问确认信息的过程，提出了基于问询历史的开放域对话系统前后文一致性评估框架。该框架通过对话机器人间对话交互取代人机对话交互，极大提高了对话数据收集效率，并在对话机器人间对话中插入与对话历史中事实或观点相关的问题，验证被测对话机器人是否可以保持前后文一致。实验证明，该框架可以有效、高效、稳定地进行开放域对话系统前后文一致性自动评估。

关键词：开放域对话，多模态对话，对话自动评估，对话历史建模

Abstract

Open-domain dialogue systems refer to meaningful dialogue in an open domain. It is an important research direction in artificial intelligence and natural language processing. In recent years, with the construction of large-scale high-quality dialogue data and the development of large-scale pre-trained language models, open-domain dialogue systems have made great progress. In open-domain human-bot dialogue interaction, the performance of open-domain dialogue systems is not satisfactory due to the high diversity of dialogue process, frequent exchange of dialogue information and long dialogue interaction process. These characteristics lead to the problems: open domain dialogue system context incoherence, general response, knowledge illusion, context inconsistency and other phenomena. Open-domain dialogue systems essentially want to be able to talk and interact like humans. In the process of dialogue cognition and learning, human beings need to fully understand the dialogue logic from the dialogue content, obtain relevant knowledge from the outside world (including visual, auditory and textual information), and keep the dialogue consistent. Accordingly, these phenomena can be attributed to the following problems related to human cognitive process of dialogue: confusion of dialogue logic, lack of relevant background knowledge and lack of consistency of dialogue.

Focused on these three problems and simulating the human cognitive process of dialogue, this paper carries out the following research:

1. **Open-domain Dialogue History Modeling with High-level Semantic Information Flow**

In open-domain human-bot interactions, due to the diversity of dialogue process and the difficulty of understanding the dialogue logic, open-domain dialogue systems often occurs the phenomenon of logic confusion, which greatly affects the human-bot interaction experience. This phenomenon is attributed to the lack of modeling ability of open-domain dialogue systems to fully understand the dialogue history. At present, there are two main ways of open-domain dialogue history modeling: flat mode and

hierarchical modeling. In the flat mode, concatenating conversation history makes it difficult for the model to capture relationships between sentences. In the hierarchical modeling, the conversation context information is missing when the conversation statements are encoded separately. Therefore, this paper proposes an open-domain dialogue history modeling method based on high-level semantic information flow to alleviate the confusion of dialogue logic. In this method, the context information is fully considered in the low-level dialogue representation and the semantic information flow between the high-level dialogue representation is modeled. Thus, the content of the future dialogue is planned according to the existing dialogue logic, and the history of the dialogue is modeled. On this basis, the DialoFlow dialogue pre-training model and the flow score are designed to demonstrate the effectiveness of the DialoFlow dialogue history modeling method introduced in this paper in the experiments of dialogue response generation and automatic dialogue evaluation.

2. **Multi-modal Knowledge Incorporated Open-domain Dialogue Response Generation**

In the process of conversation, people will acquire various modal information from the surrounding environment, such as visual information, auditory information and text information. The multi-modal information is complementary to each other, helping people understand knowledge and engage in dialogue. Conversational chatbots can produce generic or incorrect responses if they lack relevant knowledge. In order to make the open-domain dialogue system have the ability to understand multi-modal information, we incorporate the multi-modal knowledge such as video, audio, text into open-domain dialogue systems. We mainly focus on the lack of multi-modal dialogue data and the modal fusion difficulties and propose a universal multi-modal transformer with three training objectives. Experiments show that the general multi-modal transformer model proposed in this paper can effectively fuse multi-modal information in open-domain dialogue systems and generate smooth and informative dialogue responses.

3. **Automatic Evaluation for the Consistency of Chatbots with Inquiry about Dialogue History**

At present, the open-domain dialogue systems can generate fluent, diverse, infor-

mative and context-related responses in the process of human-bot interaction, but there are still many deficiencies in the consistency of the dialogue, especially within the discussion of ideas or facts, where there are often contextual conflicts. In the process of human-bot interaction, inconsistency between responses and context will greatly affect the experience of human-bot conversation. Consistency is one of the basic requirements of open-domain chatbots. An effective and efficient evaluation method is needed to improve the consistency of open-domain chatbots. However, the current evaluation methods rely heavily on manual evaluation which cannot be repeated, and there is no effective, efficient and stable automatic evaluation method. Therefore, this paper simulates the process of human repeatedly asking for confirmation information in conversation, and proposes an evaluation framework for contextual consistency of open domain conversation system based on inquiry about dialogue history. The framework replaces human-bot dialogue with bot-bot dialogue to greatly improve the efficiency of dialogue data collection, and inserts questions related to the facts or opinions in the dialogue history into the dialogue between chatbots to verify whether the chatbot can maintain the consistency of the context. Experimental results show that this framework can effectively, efficiently and stably evaluate the consistency of open-domain chatbots.

Keywords: Open-domain Dialogue, Multi-modal Dialogue, Dialogue Consistency Automatic Evaluation, Dialogue History Modeling

目 录

图形列表

表格列表

第 1 章　引言

1.1　研究背景与意义

人工智能近几年来的发展非常迅速，这主要是得益于大数据、大规模模型、大规模计算的进步和突破。根据人工智能的不同概念，可以将人工智能大致分为两个方向，感知智能和认知智能。在上个十年，感知智能中的计算机视觉进展迅速，但认知智能中的自然语言理解发展速度有限。因此，在未来的十年，人工智能的突破将会依靠于自然语言的理解的进步。语言是人类相互交流和理解的桥梁，也是人与机器沟通的自然纽带。人机对话是自然语言理解的一个重要领域，有希望成为元宇宙中主要交互方式。

自 1950 年艾伦 · 图灵提出图灵测试以后，基于自然语言的人机交互逐渐受到越来越广泛的关注。简单来说，理想的对话系统就是类似于科幻电影《钢铁侠》中的 Jawis 人工智能助手，我们期待有一天机器能够像人一样流畅、自然、富有感情、饱含人设地去和人对话，像人一样思考，帮助人完成任务，并且人类还无法辨认对方是真实的人还是机器。目前很多国内外的公司，如腾讯、阿里、小米、Facebook、Google、Amazon、Microsoft 都在投入大量的人力物力研究和打造智能对话系统，如微信小微、天猫精灵、小爱同学、Google Assistant、Amazon Alexa、微软小冰等已经融入到人们的日常生活中。

这些智能对话系统在任务型对话可以达到较好的效果，即帮助用户完成一些简单的任务，如查询天气、预定餐厅、酒店、机票等。这是由于在任务型对话中，用户的对话多样性较低，任务流程较为固定，通过收集大量的数据可以较好地提升任务型对话的能力。但是目前现有的对话系统在开放域对话（在开放的领域内进行有意义的对话，如闲聊等）上表现不尽如人意，主要表现在对话系统在与用户交互过程中常常前后不连贯或答非所问、生成通用回复和出现幻觉、前后文对话信息不一致等。开放域对话的主要难点在于多轮对话逻辑复杂多变，对话交互过程需要丰富的背景知识的支撑，对话前后文长距离需要保持一致。从人类对话认知过程讲，人类在对话交互中需要从充分理解对话逻辑、从外界获取相关背景知识、保持前后文对话一致。开放域对话系统终极目标是可以像人一样交互。所以，如何从人类认知的角度出发，分析开放域对话系统存在的问题，模

拟人类认知行为来提高开放域对话的能力，是未来开放域对话研究的热点。

1.2 研究现状

开放域对话系统目前是自然语言处理领域的一个热点研究问题，基于深度神经网络的方法成为了新的研究点。总的来说，现在的开放域对话系统研究主要分为两大类：（1）检索式开放域对话系统；（2）生成式开放域对话系统。下文将按照这两大类进行详细的介绍。

1.2.1 检索式开放域对话系统

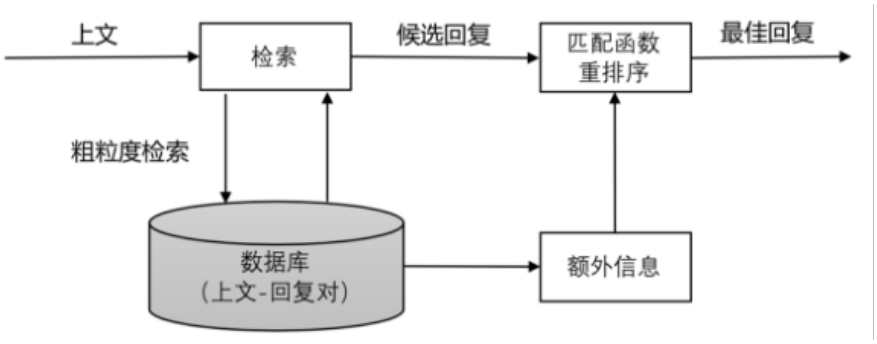

图 1.1 检索式开放域对话系统的流水线结构

Figure 1.1 The pipeline of retrieval-based open-domain dialogue systems

检索式开放域对话系统是根据对话上文从检索库中选择一个合适的回复。此类方法需要提前人工构建一个大规模的检索对话语料库，语料库的质量决定了该检索式对话系统质量的上限。由于采用检索方法的对话系统的回复较为稳定和易于控制，因此目前大多数商用对话系统都采用检索的方案。

基于检索的对话系统的流水线结构如图1.1所示，系统输入为对话上文，系统将输入的对话上文作为查询语句，按照某种检索算法（例如，BM25 算法、TF-IDF 算法）召回相关的前 k 个候选回复，然后根据匹配函数重排序算法对候选回复进行重排序，最终选择分数最高的回复作为最佳回复。基于检索的对话系统的核心模块在于匹配函数重排序算法，即如何精细化度量对话上文与回复的匹配程度，这将直接影响整个检索式对话系统的回复质量。早期检索式对话系统大多使用基于统计的方法，如 TF-IDF 算法，此类算法无论是粗粒度检索还是重排序算法都使用的是浅层的词级别信息匹配，无法做到语义层面的深度匹配。

随着深度神经网络的发展，检索式对话系统逐渐采用深度神经网络结构来表

示对话上文和回复的特征，利用特征之间的语义匹配度来选择出最合适的回复。其中常采用的神经网络结构主要有卷积神经网络（CNN）(Krizhevsky 等, 2012)，循环神经网络（RNN (Mikolov 等, 2010)，LSTM (Hochreiter 和 Schmidhuber, 1997) 和 GRU (Chung 等, 2014) 等），和基于 Transformer 的模型（BERT (Devlin 等, 2019a)，RoBERTa (Liu 等, 2020b) 等）。

从对话历史建模轮数来划分，主要分为单轮对话检索和多轮对话检索。单轮对话是指不考虑对话历史，只关注对话前一句，然后产生合适的回复。单轮对话包含的上下文信息有限，与真实的人机交互应用场景不符。因此多轮对话是目前检索式对话系统研究的主流。多轮对话是指根据多轮对话历史，产生合理的回复。多轮对话历史为回复带来了更多的信息，同时也带来了更多的约束。

从采用的方法框架划分，根据是否进行语义表示匹配交互，主要分为两大类:

1. **独立语义表示匹配方法**

独立语义表示匹配方法分别将对话上文和候选回复表示成固定长度的低维语义向量，然后利用不同的匹配函数计算二者之间的匹配分数。Hu 等 (2014) 提出使用卷积神经网络进行语句表示，使用多层感知机作为匹配函数。Lowe 等 (2015); Inaba 和 Takahashi (2016); Zhou 等 (2016); Yan 等 (2016) 等旨在改进对话上文的编码方式，从简单拼接，到层次结构，到多粒度表示，通过得到更好的表示向量与候选回复向量计算出更合理的匹配分数。独立语义表示匹配方法先编码成固定长度向量后交互的方式会导致忽略掉许多细节信息，从而难以得到最优的匹配计算。

2. **交互语义表示匹配方法**

交互语义表示匹配方法解决了独立语义表示匹配方法存在的问题。该方法充分建模了对话上文中每句话与候选回复之间的交互，在细粒度上计算了二者的匹配程度。图1.2展示了交互语义表示匹配方法的基本架构。Wu 等 (2017); Zhou 等 (2018c); Tao 等 (2019); Lu 等 (2019); Yuan 等 (2019) 主要基于传统 LSTM 架构或者 Transformer 架构，依据人类经验设计各种复杂的匹配网络去完成匹配计算。随着大规模预训练模型的兴起，大规模参数、充分的预训练为句对间的充分交互提供了强有力的支持，因此基于 LSTM 和 Transformer 的方法逐渐被大规模预训练模型所取代 (Gu 等, 2020a; Lu 等, 2020; Wang 等, 2020; Li 等, 2021a)。

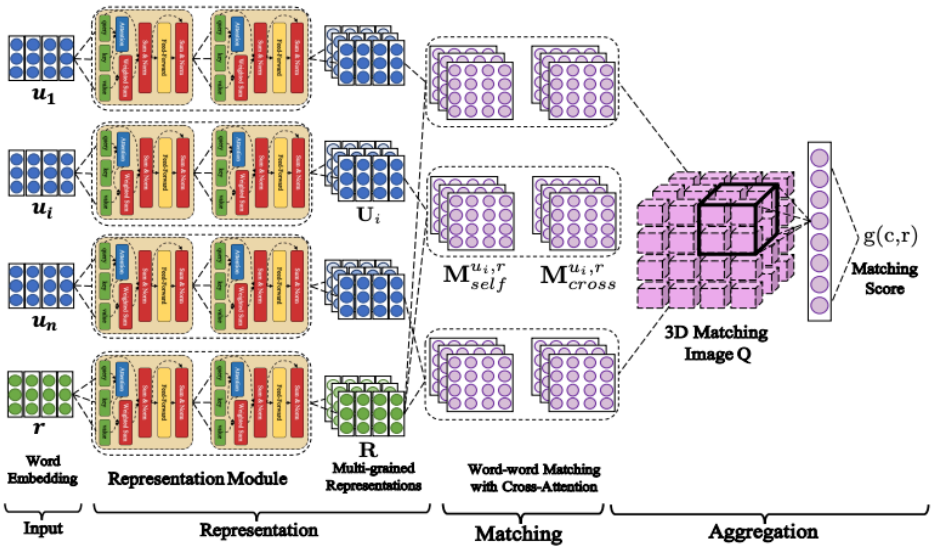

图 1.2 基于深度语义表示匹配的检索式开放域对话框架 (Zhou 等, 2018c)

Figure 1.2 The framework of retrieval-based open-domain dialogue model with deep semantic matching(Zhou 等, 2018c)

虽然这种基于检索的方法能够保证对话回复流畅，但是仍然存在很多挑战，例如方法不够灵活、应变能力不足、回复多样性不足、系统效果严重依赖人工构建的检索数据库，不易加入额外的信息等。因此，研究人员提出使用深度文本生成的方式进行对话生成。

1.2.2 生成式开放域对话系统

基于深度文本生成的对话系统更接近人类的对话方式，通过编码包含 N 句对话语句的对话历史 $\mathbf{c}_n = \{\mathbf{u}^1, \mathbf{u}^2, ..., \mathbf{u}^N\}$ 得到稠密表示向量，其中 $\mathbf{u}^i = \{u_1^i, u_2^i, ..., u_M^i\}$ 代表一个包含 M 个词的对话语句 $\mathbf{u}^i$，然后解码器生成对话回复 $\mathbf{r} = \{r_1, r_2, ..., r_K\}$，其中 K 是对话回复的长度。基于深度神经网络的对话生成模型通过优化在给定对话历史的情况下的标准对话回复的概率进行训练：

$$P(\mathbf{r}|\mathbf{c}) = \prod_{k=1}^{K} P(r_k|\mathbf{r}_{<k}, \mathbf{c}; \theta) \tag{1.1}$$

其中 θ 为模型参数。

本小节主要介绍目前被广泛应用的几种最主要的开放域对话生成模型：

1. 基于注意力机制的编码器-解码器模型

2015 年，Bahdanau 等 (2014) 首次在神经机器翻译中提出基于注意力机制的 Seq2Seq 模型，Sordoni 等 (2015) 和Shang 等 (2015) 开始将基于注意力机制的

Seq2Seq 模型用于对话生成。如图 1.3所示。该模型基于门控循环单元（Gated Recurrent Unit, GRU）。它首先将对话历史中的词语 x_i 通过词嵌入层编码成稠密

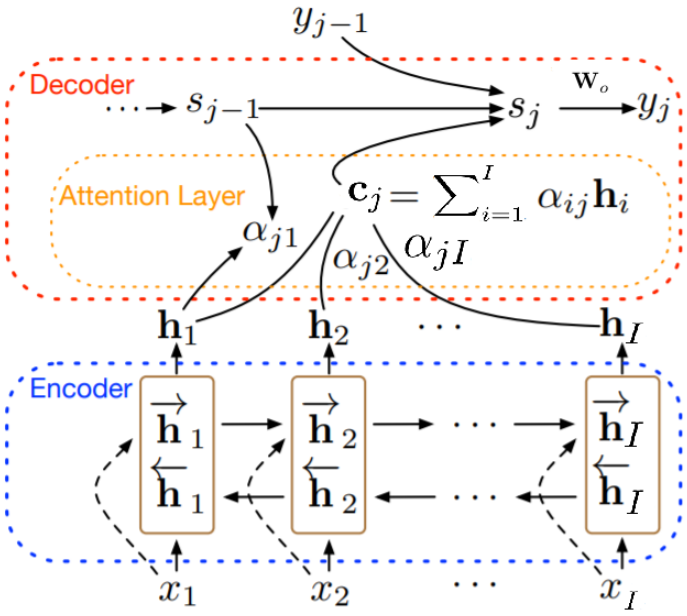

图 1.3 基于注意力机制的编码器-解码器的深度生成模型结构图 (Zhang 等, 2018b)

Figure 1.3 The architecture of attention-based seq2seq model(Zhang 等, 2018b)

的嵌入 $\mathbf{x}_i$，然后通过一层 GRU 进行向量映射，得到一个固定维度的向量表示 $\mathbf{h}_i$：

$$\overrightarrow{\mathbf{h}_i} = \overrightarrow{\text{GRU}}\left(\mathbf{x}_i, \overrightarrow{\mathbf{h}}_{i-1}\right)$$

$$\overleftarrow{\mathbf{h}_i} = \overleftarrow{\text{GRU}}\left(\mathbf{x}_i, \overleftarrow{\mathbf{h}}_{i+1}\right)$$

$$\mathbf{h}_i = [\overrightarrow{\mathbf{h}_i}; \overleftarrow{\mathbf{h}_i}]$$

解码器由一层门控循环单元和一个注意力模块组成。解码的目标是最大化目标语言句子的生成概率。在生成目标句子的每一个词语时，利用注意力机制计算它对应输入句子的注意力应该在哪些词语上。注意力机制使用一个多层感知机计算解码器中的隐状态 $\mathbf{s}_j$ 与对话历史中每个词 $\mathbf{x}_i$ 的稠密语义向量 $\mathbf{h}_i$ 的相关性权重 $\boldsymbol{\alpha}_{ij}$，然后根据该权重对 $\mathbf{h}_i$ 进行加权求和，得到注意力机制的输出 $\mathbf{c}_j$：

$$e_{ij} = \mathbf{v}_a^T \tanh(\mathbf{W}_a \mathbf{s}_{j-1} + \mathbf{U}_a \mathbf{h}_i)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{i'=1}^{I} \exp(e_{i'j})}$$

$$\mathbf{c}_j = \sum_{i=1}^{I} \alpha_{ij} \mathbf{h}_i$$

其中 $\mathbf{v}_a \in \mathbb{R}^d$ 、$\mathbf{W}_a \in \mathbb{R}^{d\times d}$ 和 $\mathbf{U}_a \in \mathbb{R}^{d\times 2d}$ 表示权重矩阵，d 表示隐藏层单元大小。随后，解码器的门控循环单元将解码器中隐状态的 $\mathbf{s}_{j-1}$、对话回复输入 $\mathbf{y}_{j-1}$、注意力机制的输出 $\mathbf{c}_j$ 作为输入并生成当前的预测：

$$\mathbf{s}_j = f(\mathbf{s}_{j-1}, \mathbf{y}_{j-1}, \mathbf{c}_j)$$

$$p(y_j|\mathbf{y}_{<j}, \mathbf{x}) \propto \exp(f(\mathbf{s}_j, \mathbf{y}_{j-1}, \mathbf{c}_j) \cdot \mathbf{W}_o)$$

其中，$\mathbf{W}_o$ 表示解码器输出层的参数，f 是一个非线性映射函数。虽然基于注意力机制的 Seq2Seq 模型在对话回复生成中表现出不错的效果。然而，此类基于注意力机制的 RNN 结构的序列到序列模型只支持串行训练，难以用于大规模快速并行训练。

2. **基于多层次注意力的编码器-解码器模型**

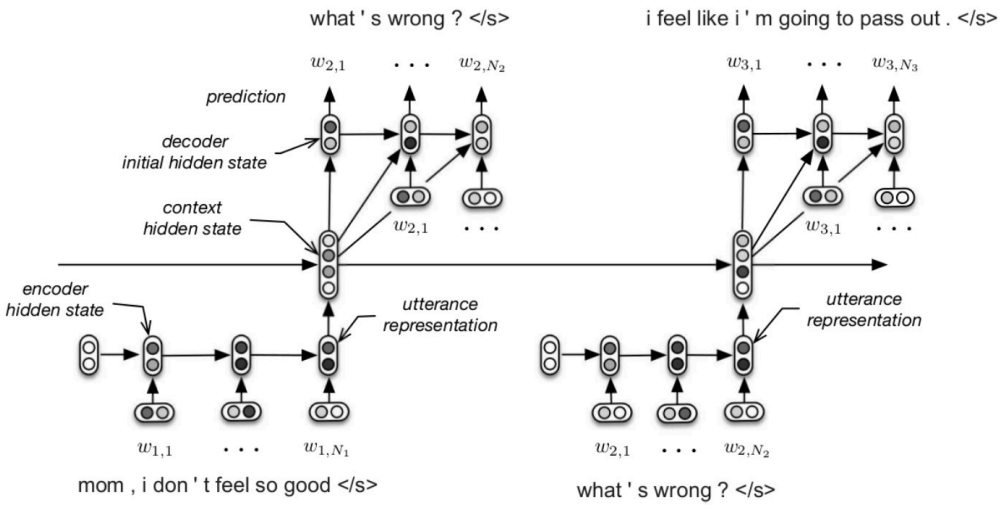

图 1.4 多层次注意力的编码器-解码器模型架构图 (Serban 等, 2016b)

Figure 1.4 The architecture of hierarchical encoder-decoder model(Serban 等, 2016b)

对话系统一般拥有多轮对话历史，早期Sordoni 等 (2015) 将多轮对话上文进行拼接得到上文历史信息的编码表示，然后通过解码器进行一个词一个词的输出。由于多轮对话历史信息较长，通过拼接的方式在一定程度上忽略了句子间的语序关系。随后Serban 等 (2016b) 提出使用层次化的 RNN 编码器-解码器（Hierarchical Encoder-Decoder, HRED）解决多轮对话的生成问题，如图1.4所示。

该模型首先在词语级别使用词级别 RNN 进行句子的表示，然后在高层使用另一个句级别 RNN 对上文历史句子的表示进行再编码，得到对话级别的向量表示，通过解码器使用该向量表示进行回复的生成。随着 HRED 在多轮对话中的

广泛应用，越来越多的变体模型被提出。Serban 等 (2016c, 2017) 提出了使用包含隐变量作为中间状态的 VHRED 和 Mr-RNN，他们认为生成回复应该进行初步的分类，通过对隐变量加上高斯分布的约束，可以获得更多样更鲁棒的文本生成。随后，Tian 等 (2017); Xing 等 (2017) 借鉴传统的注意力机制 Seq2Seq 模型 (Bahdanau 等, 2014)，将这种注意力机制引入到 HRED 结构中。随着自注意力模型 Tranformer 和大规模预训练语言模型在自然语言处理领域的快速发展，Shan 等 (2020); Gu 等 (2021) 提出了使用预训练 BERT 模型编码单句对话，再采用句子级别编码器捕捉对话整体结构，大大提升了对话历史建模的效果。

3. **基于完全注意力的 Transformer 模型**

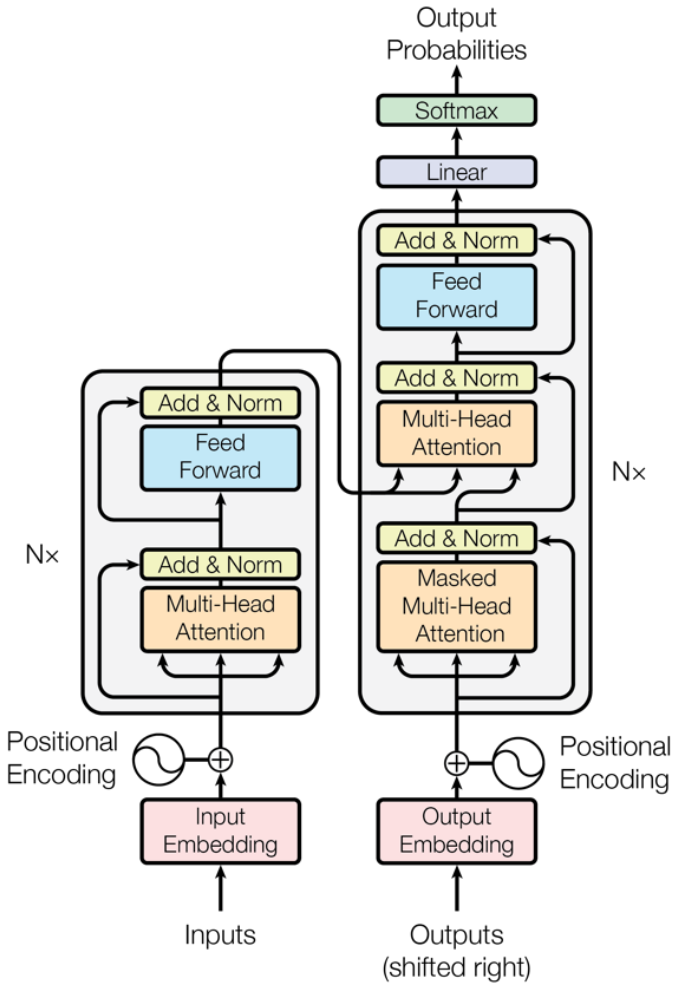

图 1.5 基于完全注意力的 Transformer 模型架构 (Vaswani 等, 2017b)

Figure 1.5 The architecture of transformer(Vaswani 等, 2017b)

谷歌大脑于 2017 年新提出了一个基于完全注意力机制的、可以并行训练的、用于序列到序列生成任务的新型模型 Transformer (Vaswani 等, 2017a)，模型架构

如图1.5所示，该架构由堆叠了 N 个基本块的编码器和解码器组成。对于编码器来说，每块均由多头注意力（Multi-Head Attention）子层和前馈神经网络（Feed Forward）子层构成，子层之后都进行了残差（Residual Add）和层正规化（Layer Normalization）操作。对于解码器，其整体结构与编码器保持不变，只是在多头注意力子层和前馈神经网络子层之间引入了用于与源端进行交互的编码器-解码器多头注意力子层。在训练过程中，为了防止探测到未来的信息，解码器的多头自注意力子层是掩码表示的（Masked Multi-Head Attention）。由于模型只使用了注意力机制，并不具备像循环神经网络一样对语言位置信息建模的能力，因此模型引入了位置编码（Positional Encoding），使用基于周期性的正弦变换和余弦变换对句子中单词的位置信息进行编码，使模型拥有对位置信息建模的能力：

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

其中 pos 为当前词的绝对位置，i 为对应的词嵌入所在的维度，d_{model} 是模型的维度。稠密的词嵌入加上相应位置的位置编码构成了模型的输入。

多头注意力模块是 Transformer 模型结构中比较重要的模块，结构如图1.6所示：对于给定的查询 **Q**、键 **K**，值 **V**，多头注意力模块首先通过不同的线性变

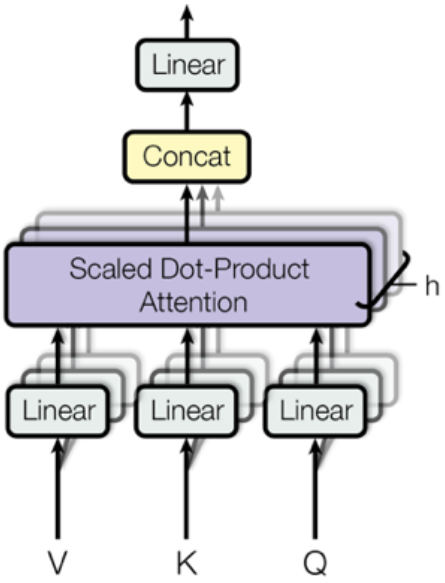

图 1.6 多头注意力 (Vaswani 等, 2017b)

Figure 1.6 The multi-head attention (Vaswani 等, 2017b)

换层将他们映射到 h 个不同的子空间，分别对映射之后的稠密表示计算注意力，

再将不同空间中的注意力表示拼接起来，最后再对拼接后的注意力进行再次映射得到最终输出：

$$\mathbf{q} = \mathbf{Q}\mathbf{W}_Q = [\mathbf{q}_1; \mathbf{q}_2; \cdots; \mathbf{q}_h]$$
$$\mathbf{k} = \mathbf{K}\mathbf{W}_K = [\mathbf{k}_1; \mathbf{k}_2; \cdots; \mathbf{k}_h]$$
$$\mathbf{v} = \mathbf{V}\mathbf{W}_V = [\mathbf{v}_1; \mathbf{v}_2; \cdots; \mathbf{v}_h]$$
$$\mathbf{o}_i = \text{Attention}(\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i)$$
$$\mathbf{o} = [\mathbf{o}_1; \mathbf{o}_2; \cdots; \mathbf{o}_h]\mathbf{W}_O$$

其中 $\mathbf{W}_Q$、$\mathbf{W}_K$、$\mathbf{W}_V$、$\mathbf{W}_O$ 均为线性变换的权重矩阵，Attention 为放缩点积注意力（Scaled Dot-Product Attrntion）：

$$\text{Attention}(\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i) = \text{softmax}(\frac{\mathbf{q}_i \mathbf{k}_i^T}{\sqrt{d_{model}}})\mathbf{v}_i \tag{1.2}$$

Transformer 支持并行训练这一优良品质，为大规模预训练模型的出现提供了天然基础。预训练模型也给对话生成任务带来了很大的性能提升。因此，对于对话生成而言，目前主流的建模方法是基于预训练模型（例如，GPT2 (Brown 等, 2020a)，T5 (Raffel 等, 2020)）把整个对话历史拼接成一个线性序列的文本作为编码器输入，然后通过自注意力机制得到每个词语的向量表示，然后同上述 RNN 结构的 Seq2Seq 架构一样，解码器逐步生成出每个词语，近期主要工作包括Zhao 等 (2020); Dinan 等 (2020a); Bao 等 (2020b); Liu 等 (2020a) 等。

1.2.3　主要研究趋势

开放域对话系统的主要功能就是与人进行交互，进行有意义的对话。因此，如何提升开放域对话系统的人机交互体验，提高与用户的交互时间，完成用户长期和复杂的目标是非常重要的研究课题。当前，开放域对话的前沿探究大致分为以下几个方向：

1. 发挥大规模对话预训练模型的能力

随着大规模预训练语言模型技术的发展，在对话生成领域，针对对话建模的大规模预训练语言模型，如 DialoGPT (Zhang 等, 2020a)、Plato-2 (Bao 等, 2020a)、Meena (Adiwardana 等, 2020a)、Blender (Smith 等, 2020a)、DialoFlow (Li 等, 2021e) 等，开始在各种下游任务中大放异彩。预训练模型的发展已经使得一些基础的问

题得以解决，例如信息量、多样性等，也为很多任务提供了新的基准。预训练技术已经成为了整个领域发展的里程碑，也将是未来开放域对话系统系列任务的基础，例如回复生成、角色化、风格化、推理等。

2. **提升开放域对话系统的知识丰富性**

对话系统需要了解人类世界的知识，才能在与用户对话交互过程中将对话进行下去。对话具有信息量，使得用户可以在与对话系统交互过程中获取信息，这也是开放域对话系统的意义所在。目前在开放域对话系统中引入知识的研究热点主要集中于在不同场景下引入多源异质的多模态信息，如在元宇宙的应用场景中，开放域对话系统需要对周围的环境进行感知，包含文本信息、视觉信息、听觉信息等。目前有很多研究工作，如Young 等 (2018); Liu 等 (2018); Zhou 等 (2018a) 通过知识图谱的方式将人类世界的常识知识引入到对话系统中。Ghazvininejad 等 (2018); Han 等 (2015) 建立了文档知识库，探索让对话系统理解非结构化文本知识，辅助对话生成。Li 等 (2021b); Fei 等 (2021) 尝试将多模态信息，如视频、音频、表情等引入对话系统。

3. **提升开放域对话系统的前后文一致性**

一个好的对话系统需要保持前后文一致性，这样才能获得用户的信任和好感。前后文一致性主要包含角色一致性与对话历史一致性。目前的对话系统的角色化一致性研究主要有两种方法：隐式角色化和显式角色化。前者是将角色信息编码成一个隐式向量 (Li 等, 2016; Yang 等, 2017; Wang 等, 2017; Zhang 等, 2019b)；后者一般是提供一个角色信息表（例如姓名、年龄、性别、星座、爱好等），对话系统首先显式选择有用的角色化信息，然后参与解码生成回复 (Qian 等, 2018; Zhang 等, 2018a; Zheng 等, 2019)。对于前后文一致性，早期的工作主要是更好的建模对话上下文使得上下文隐式地保持一致，直到Welleck 等 (2019a); Song 等 (2020a) 将检测上下文一致性建模成自然语言推理（NLI）问题。2021 年初，Nie 等 (2020a) 人工标注了一个小规模的对话冲突检测数据集用于促进这一研究方向的发展。

4. **优化开放域对话质量自动评估方法**

开放域对话质量自动评估是希望设计一种有效、可靠、高效的评估方法，判断对话系统生成的回复在对话的各个维度上的好坏，与人类认知高度相关。目前大多数研究工作使用的开放域对话质量自动评估方法为 BLEU、ROUGE 等机器

翻译中常用的评价翻译忠实度的方法。而开放域对话具有一对多的特性，即同一个上文对应多个可能的回复，因此，此类基于评价生成的回复与参考回复的相似度的方法并不适用于对话生成的任务。缺乏有效高效的开放域对话质量自动评估方法限制了开放域对话系统能力的迭代和提升。目前一些研究工作开始着眼于使用基于模型的方法进行对话质量自动评估，此类方法无需给出参考回复，故有较强的适用性。Mehri 和 Eskénazi (2020a) 提出训练多种模型评估对话的不同维度，如流畅度、上下文相关性、知识性等。Gao 等 (2020) 集成了多种 GPT2 模型，根据对话系统用户反馈进行训练，以此来判断对话回复的类人性。Li 等 (2021d) 提出了一种开放域对话前后一致性评估框架。开放域对话评估的最大的挑战是如何设计出与人类认知高度相关的评估方法，需要解决以下问题：(1) 在人类认知中，哪些对话维度比较重要；(2) 针对某个对话维度，如何设计评估指标真实模拟人类在此维度的认知习惯。

1.3　人机对话交互中的认知过程

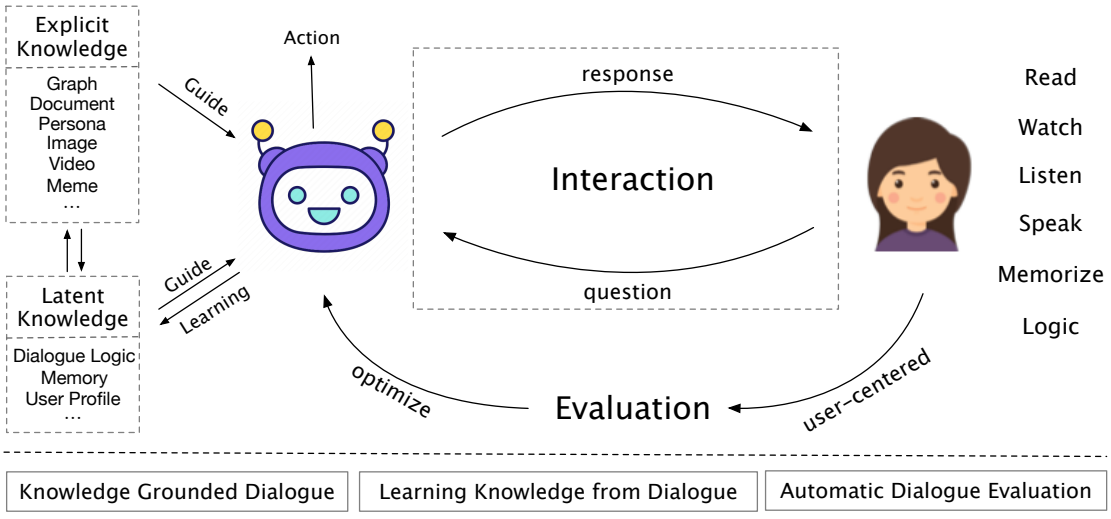

图 1.7 人类对话交互中的认知过程

Figure 1.7 The cognitive process of conversation for human beings

图1.7展示了人机交互中的认知过程。人类在对话过程中，有看、听、读等动作，需要从外界获取多种多样的知识，如视觉指示、听觉知识、文本知识，这些知识相辅相成，协助人类进行对话的理解与对话回复。人们在对话交互过程中也会利用自身的记忆，理解当前对话逻辑，回复恰当的话来达到自己的对话目的。开放域对话系统本质上就是要像人一样进行对话交互，所以开放域对话系统也

应该从外部获取多模态知识以及内部记忆，理解对话内容，按照人类的对话逻辑，生成恰当的回复。同时，参考人们在学习对话时会得到其他人对对话好坏的评价，开放域对话系统评估也应该以人为中心，使用人们期望中的对话特性或指标进行评价。

1.4 主要研究难点

尽管开放域对话系统在多种场景、多个数据集上的表现有很大进步，但是现有的开放域对话系统对对话问题的建模仍然存在很多缺陷。本文主要从小节1.3中提到的人类对话认知过程出发，针对开放域对话中的以下几个问题进行研究：

1.4.1 开放域对话系统对话逻辑混乱问题

开放域对话系统通过理解对话历史，生成合适的对话回复。但是在实际应用过程中，开放域对话系统经常会出现与上文无关、前言不搭后语的现象，统称为逻辑混乱问题。例如，对话历史为“我刚看了《复仇者联盟4》，非常喜欢电影中的钢铁侠”，对话回复为“我现在还单身，没有女朋友”。此类对话过程会让用户感觉摸不到头脑，会极大影响人们的对话交互体验。此类问题的根源是开放域对话系统对对对话历史的理解能力不足，没有从对话语料中学习到正确的对话逻辑。目前随着大规模对话语料的构建和大规模模型的发展，开放域对话系统可以较好的处理单轮对话，但是对于多轮对话的前后语义关系理解能力不足，导致无法很好地捕捉到整体的对话逻辑，导致对话回复的逻辑混乱问题。解决开放域对话系统对话逻辑混乱问题首先需要对对话历史进行有效的建模，从对话语料中学到正确的对话逻辑。

1.4.2 开放域对话系统缺乏背景知识问题

在人机对话过程中，用户经常聊到一些与背景知识相关的信息，例如用户早上看到的新闻或视频，想与对话系统根据这些话题进行聊天。如果对话系统不具备这些知识，那么就很容易生成通用的回复或生成包含错误知识的回复，例如：“用户：你看过《复仇者联盟4》吗？对话系统：我没有看过。”或“用户：你看过《复仇者联盟4》吗？对话系统：讲的是复仇者帮助灭霸消灭人类的故事。”这些现象归结于对话系统缺乏背景知识或没有有效利用背景知识，会导致对话过程

结束和用户满意度降低。人们在日常对话交互中会从周围环境中获取多种多样的知识，同样开放域对话系统中的背景知识也是多源异质的，大体包含以下几类：文本（例如新闻、文档段落等），图片（例如表情、微博配图等），视频（例如电影、短视频等），音频（例如音乐、自然声音等）。多源异质知识由于来源以及组织方式不同，从而需要不同的表示方式。在该问题中，存在知识标注的对话数据较少，多种模态与来源的知识表示与融合难度较大。因此，如何使对话系统有效利用到这些知识，生成流畅、正确、有信息量的回复是一个重要的研究问题。

1.4.3　开放域对话系统前后文一致性自动评估难问题

在人类对话交互过程中，前后一致性是非常重要的特性，前后文不一致会极大影响对话交互的体验。目前，开放域对话系统在人机交互过程中可以生成出流畅的、多样的、有信息量的、与对话历史相关的回复，但是在对话前后文一致性上还存在很多不足之处，尤其在讨论关于观点或事实时经常产生前后文冲突的现象。开放域对话系统前后文一致性的提高离不开有效的评估方式，目前开放域对话系统前后一致性目前最有效的评估方式是通过人工评估。首先通过人工与对话系统进行对话交互，再通过人工对整段对话进行评估得到对话系统的前后文一致性。然而，人工评估耗时耗力、成本过高、扩展性差，并且与参与评估的人工评估者的认知能力强相关。缺乏一种有效且高效的开放域对话系统前后一致性自动化评估方法限制了开放域对话生成能力的迭代与提升。

1.5　主要研究内容

本文主要从人类对话认知过程出发，针对上文中提出的开放域对话系统存在的难点和挑战，提出了相应的解决方法。具体研究工作包括：

1.5.1　引入高层语义信息流的对话历史建模

本文提出引入高层语义信息流进行对话历史建模来缓解对话逻辑混乱问题。为了同时在低层对话语句编码中充分考虑上下文信息与在高层建模对话语句间的关系，受人类认知过程“人类总是在继续对话之前考虑下一个对话回复对整体对话的影响 (Brown-Schmidt 和 Konopka, 2015)”启发，本文提出了 DialFlow 模型，引入了一个 Flow 模块，建模对话高层语义信息流，预测未来对话的发展方向，指导对话回复生成。通过这种方式，模型可以对对话历史有更全面的理解，

也捕捉到整体对话的逻辑与对话未来发展方向，从而缓解对话逻辑混乱的问题。

1.5.2 引入多模态知识的开放域对话回复生成

本文提出了一个统一的多模态 Transformer 模型来将多模态知识引入对话系统中。该模型针对多模态对话数据稀缺的问题，在预训练语言模型的基础上增加多模态输入，然后设计了多种训练任务促进多模态知识的表示与融合。具体地，本文引入了视听场景感知信息，设计了视频音频对齐和嵌入方式，将视频音频的表示转换到文本的表示空间，同时基于大规模预训练语言模型进行回复生成建模、视频音频序列建模、视频描述建模三种任务的学习，以此促进视频音频信息与文本信息的融合表示。通过这种方法，一方面降低了视听场景感知的对话对数据量的需求，另一方面可以有效地在对话生成中引入多模态知识。

1.5.3 引入问询的开放域对话系统前后一致性自动评估研究

本文针对开放域对话交互中前后一致性人工评估耗时耗力、成本过高、扩展性差问题，提出了一种引入问询的开放域对话系统前后一致性评估框架。该框架基于聊天机器人与聊天机器人之间的对话，相比于人机交互大大节省了时间与成本。同时，该框架针对开放域对话交互中容易产生关于事实和观点前后不一致的问题，模拟人类在对话交互过程中通过询问确认信息的行为，设计了问询模式，通过一个问题生成器针对对话历史进行发问，检测被测聊天机器人是否可以正确回答该问题。通过这种方式，该框架可以大大节省时间和成本，也可以产生有效、稳定的评估结果。

1.6 章节组织

本文的组织结构如下：

第一章介绍了开放域对话系统的研究背景、意义和发展现状，介绍了开放域对话系统的基本方法、主要模型和主要研究方向，并介绍了本文的主要研究内容。

第二章主要模拟人类知识过程，建模对话历史中的高层次对话信息流，提出了基于高层次信息流建模的开放域对话模型，在大规模对话语料上进行了预训练，可以大大提升模型对对话历史的理解能力和回复生成能力。基于预训练模型，提出了无参考对话自动评估指标 Flow 分数，极大提升了开放域对话交互式

自动评估的效果。

第三章主要在开放域对话系统中引入了多模态信息，进行基于视听场景感知的开放域对话回复生成研究，设计了一个通用多模态 Transformer 模型和多个多模态融合训练任务，使得模型可以利用来自视觉、听觉、文本的信息，生成流畅、有信息量的对话回复。

第四章主要针对开放域对话机器人在交互聊天中出现的上下文不一致问题，提出了基于问询对话历史的开放域对话机器人上下文一致性评估框架。该框架可以有效、高效、稳定的评估开放域对话机器人的上下文一致性。

第五章总结了本文的所有工作，并对未来的研究方向进行了展望。

第 2 章　引入高层语义信息流的对话历史建模研究

2.1　引言

由于大规模预训练方法的快速发展 (Devlin 等, 2019b; Radford 等, 2019; Brown 等, 2020b) 与大规模高质量对话数据集的构建 (Dinan 等, 2019; Baumgartner 等, 2020; Smith 等, 2020b)，开放域对话机器人 (Adiwardana 等, 2020b; Bao 等, 2020c; Smith 等, 2020b) 取得了实质性进展。然而，开放域对话系统仍然容易出现上下文无关、前言不搭后语等对话逻辑混乱的现象。这种现象产生的原因是因为开放域对话系统不能完全理解对话历史中的逻辑，在多轮对话场景下尤为严重。要使开放域对话系统充分理解对话历史的逻辑，就需要开放域对话模型对对话历史进行充分的建模。

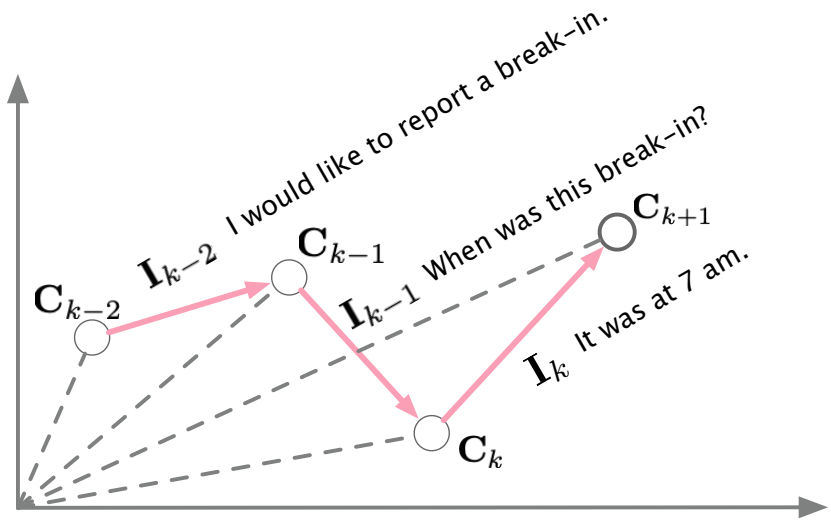

图 2.1 对话历史中的动态语义变化信息流

Figure 2.1 Semantic flow across the dialogue history

以往的对话历史建模工作主要分为两类。一类工作通常将对话历史串联起来作为模型输入并生成对话回复 (Zhang 等, 2020b; Smith 等, 2020b; Bao 等, 2020c)，本文称之为平铺模式。这种方式在大规模的预训练中被普遍采用。然而，Sankar 等 (2019) 证明，平铺拼接很可能会导致模型忽略对话历史中语句之间的动态语义变化。因此，另一类工作采用分层建模来编码对话历史 (Serban 等, 2016a; Shan 等, 2020; Gu 等, 2020b)。首先使用一个词级编码器对对话历史中的每个语句单独编码，然后将编码后的句子表示输入到一个句子级编码器，从而得到整个对话的

表示。这些方式在编码每个对话语句时缺乏对话历史信息，而对话历史信息是理解对话语句的重要条件。因此，上述两种方式在对对话历史的动态信息建模方面都存在不足。

本研究受人类认知过程的启发，“人类总是在继续对话之前考虑下一个对话回复对整体对话的影响 (Brown-Schmidt 和 Konopka, 2015)”，提出了通过分析每个对话语句所带来的语义影响来建模对话历史中的动态信息流的方法。如图2.1所示，本研究将不同话语的对话历史的稠密表示定义为语境（灰点线），将语境转换定义为每个对话语句带来的语义影响（粉色线）。本研究构建了对话语句级对话历史语义流的过程。相应地，每个对话语句带来的语义影响可以通过两个相邻语义之间的差异来衡量，来指导当前的对话回复的生成。

2.2 相关工作

本节将对开放域对话系统中与本章内容相关的研究工作进行介绍。与本章内容最相关的有三类工作：多轮对话建模；对话生成中的预训练模型；交互式对话评估方法。

2.2.1 多轮对话建模

多轮对话历史建模方法主要分为两类：（1）平铺拼接。这些工作直接将对话历史拼接为整个输入序列输入到模型中 (Zhang 等, 2020b)。模型处理的基本单位为词，而难以捕捉对话历史语句间的动态信息。（2）层次结构。层次结构是对话历史建模中常用的一种结构形式。Serban 等 (2016c) 提出了层次 LSTM 来编码对话历史和生成对话回复。Li 等 (2019a) 引入了一种增量式 Transformer 结构来捕获对话历史多轮依赖性。Shan 等 (2020); Gu 等 (2020b) 采用预训练的 BERT 模型对单个对话语句进行编码，并设计对话语句级编码器来捕获对话结构。此类方法在对对话语句进行编码时，忽略了对话历史中的词级信息。而在人类认知过程中，当前对话语句的理解与对话历史密切相关。与这些方法不同的是，DialoFlow 模型充分利用了对话历史的词级信息和对话语句级动态信息流。

2.2.2 对话生成中的预训练模型

近年来，预训练语言模型在对话生成方面取得了很大的成功。DialoGPT (Zhang 等, 2020b)、PLATO-2 (Bao 等, 2020c)、Meena (Adiwardana 等, 2020b) 和 Blender (Smith

等, 2020b) 通过在开放域对话数据上训练基于 Transformer 的语言模型，实现了强大的生成性能。相比之下，本研究提出的 DialoFlow 模型着重于对训练过程中的对话历史的动态信息流进行建模，并且设计了三个训练目标来优化模型。

2.2.3　交互式对话评估方法

自动评估交互式对话的质量是一个具有挑战性的问题。因为开放域对话没有确定性的回复，在人机开放域对话交互过程中，对话系统没有标准回复，故如何在无参考情况下进行对话质量自动评估是一个重要的研究课题。一般来说，大多数工作采用人工评估的方法进行交互式对话质量评估。最近，Mehri 和 Eskénazi (2020b) 提出了 FED 评分，一种基于预训练的 DialoGPT-large 模型的自动对话质量评估指标。它使用预先设置的常见人类评论在大规模语言模型中的困惑度，比如 “与你交谈很有趣”，来反应了对话的质量。然而，FED 的评分在那些没有明显的评论情况下的对话上表现不如人意。本研究设计的流评分完全依赖于预训练的 DialoFlow 模型，不需要预先设置人工回复，因此泛化性更好。

2.3　基于高层语义信息流的对话预训练模型

本节将描述 DialoFlow 模型的详细设计，首先介绍模型的总体设计，然后介绍详细的模型结构和训练目标，最后介绍基于该模型的流评分。

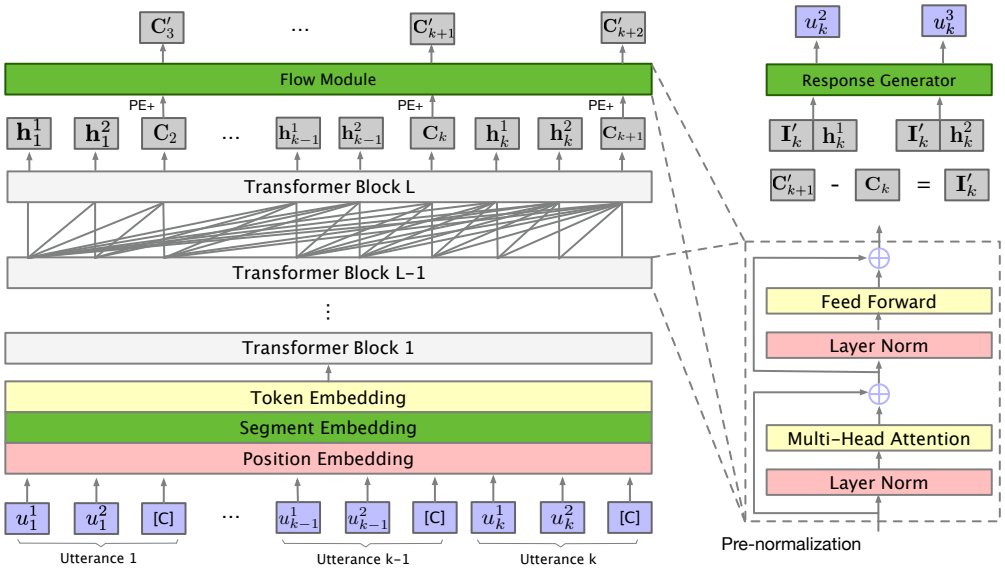

图 2.2 DialoFlow 模型结构

Figure 2.2 The architecture of DialoFlow

2.3.1 模型总览

在详细介绍 DialoFlow 模型之前，这里首先定义一些术语。形式上，令 $D = \{u_1, u_2, ..., u_N\}$ 代表整个对话。对于每个对话语句 $u_k = \{u_k^1, u_k^2, ..., u_k^T\}$，其中 u_k^t 表示第 k 句对话中的第 t 个单词。这里进一步表示 $u_{<k} = \{u_1, u_2, ..., u_{k-1}\}$ 作为第 k 句对话时的对话历史。此外，在第 k 句对话时，对话历史的稠密表示 $u_{<k}$ 被表示为语境 $\mathbf{C}_k$。第 $k+1$ 句对话的新语境 $\mathbf{C}_{k+1}$ 与第 k 句对话的前语境 $\mathbf{C}_k$ 之间的差异可以定义为第 k 句对话带来的语义影响 $\mathbf{I}_k$，可以表示为:

$$\mathbf{I}_k = \mathbf{C}_{k+1} - \mathbf{C}_k. \tag{2.1}$$

DialoFlow 模型首先编码对话历史，并根据所有之前的对话历史语境 $\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k$ 预测未来的对话语境 $\mathbf{C}'_{k+1}$。然后在对话回复生成阶段，该模型计算预测得到的目标语义影响 $\mathbf{I}'_k$，并使用预测的语义影响以及对话历史指导对话回复生成。具体来说，如图2.2所示，DialoFlow 模型通过在基础 GPT-2 模型上设计一个单向流模块对对话历史信息流进行建模，并引入三个多任务训练目标来建模对话历史信息流、对话语句语义影响和对话回复生成。

2.3.2 模型结构

图2.2展示了 DialoFlow 模型的详细结构，它由输入嵌入层、Transformer 层、单向流模块和对话回复生成器组成。

1. *输入嵌入层*

DialoFlow 模型将词嵌入、段嵌入和位置嵌入的总和作为模型输入。特别地，该模型在每句话的末尾插入一个特殊的符号 “[C]”，用来捕捉对话历史的整体稠密表示。为了增强不同说话人的建模，该模型增加了两种类型的段嵌入: “[Speaker1]” 和 “[Speaker2]”。

2. *Transformer 层*

Transformer 层由以下几个关键部件组成：正则化层、多头注意力层和前馈层。本研究使用 GPT-2 中使用的前置正则化层 (Radford 等, 2019)，而不是使用于 BERT 的后置正则化层 (Devlin 等, 2019b)。因为Shoeybi 等 (2019) 表明，post-normalization 随着模型的大小增加会导致训练不稳定及性能下降，pre-normalization 可以做到稳定的大规模模型训练。该 Transformer 层使用单向对话编码，即当前

对话语句可以看到所有对话历史，而看不到未来的对话语句。DialogFlow 模型在整体对话级别进行训练而不是对话历史-回复对上进行训练。这里可以获得由 Transformer 层编码的第 k 个对话语句的语境表示:

$$\mathbf{C}_k = \text{Transformer}(u_{<k}), \tag{2.2}$$

其中 C_k 为特殊标记 “[C]” 位置的隐藏状态。输入序列中每个标记 u_k^t 位置处的隐藏状态记为 h_k^t。

3. 单向流模块

为了捕捉对话话语之间的动态信息流，本研究设计了一个流模块来建模对话上下文变化。流模块的架构与一层 Transformer 层相同。流模块接受所有之前的上下文 $\{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k\}$ 作为输入，预测第 (k+1) 句 $\mathbf{C}'_{k+1}$ 的上下文:

$$\mathbf{C}'_{k+1} = \text{Flow}(\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k). \tag{2.3}$$

第 k 个话语所预测的语义影响可以计算为:

$$\mathbf{I}'_k = \mathbf{C}'_{k+1} - \mathbf{C}_k. \tag{2.4}$$

4. 对话回复生成器

DialoFlow 是在预测语义影响的指导下生成对话回复 u_k 的。对话回复生成器包含一个前反馈层和一个 softmax 层，用于将隐藏状态转换为词。当生成第 t 个单词时，响应生成器将隐状态 h_{t-1} 作为输入，输出第 t 个单词的概率分布:

$$\begin{aligned} &p(u_k^t | \mathbf{I}'_k, u_{<k}, u_k^{<t}) = \\ &\qquad \text{softmax}(W_1[\mathbf{I}'_k; \mathbf{h}_k^{t-1}] + b_1) \in \mathbb{R}^{|V|}, \end{aligned} \tag{2.5}$$

2.3.3　训练目标

与传统的对话历史-回复对训练方法不同，DialoFlow 模型是用包含 N 句对话的整个对话进行训练。因此，我们设计了三个训练任务来优化模型：1）对话历史语义流建模；2）对话语句语义影响建模；3）对话回复生成建模。

1. 对话历史语义流建模

为了捕获动态对话历史语义流，DialoFlow 根据前面的上下文序列 $\{\mathbf{C}_1, ..., \mathbf{C}_{k-1}\}$ 预测第 k 句对话 $\mathbf{C}'_k$ 处的语境。接下来，将预测出的对话语境 $\mathbf{C}'_k$ 和真实的对话

语境 $\mathbf{C}_k$ 之间的 L2 距离最小化：

$$\mathcal{L}_{CFM} = \sum_{k=1}^{N} ||\mathbf{C}_k - \mathbf{C}'_k||_2^2. \tag{2.6}$$

2. 对话语句语义影响建模

为了有效地对语境 $\mathbf{C}_{n-1}$ 中第 n 句对话所带来的语义影响进行建模，本研究使用预测的语义影响 $\mathbf{I}'_n$ 设计了一个词袋损失函数：

$$\begin{aligned}\mathcal{L}_{SIM} &= -\sum_{k=1}^{N}\sum_{t=1}^{T} \log p(u_k^t|\mathbf{I}'_k) \\ &= -\sum_{k=1}^{N}\sum_{t=1}^{T} \log f_{u_k^t},\end{aligned} \tag{2.7}$$

其中 $f_{u_k^t}$ 表示第 t 个单词 u_k^t 在 u_k 中出现的估计概率。函数 f 用于同时预测 u_k 话语中的所有单词：

$$f = \text{softmax}(W_2\mathbf{I}'_k + b_2) \in \mathbb{R}^{|V|}, \tag{2.8}$$

其中 $|V|$ 为词汇量，W_2 和 b_2 为可学习参数。

3. 对话回复生成建模

预测的对话语句语义影响 $\mathbf{I}'_k$ 也可以看作是第 k 句对话的语义期望。本研究将预测的语义影响 $\mathbf{I}'_k$ 引入对话回复生成阶段，以指导对话回复生成。对话回复生成目标如下：

$$\begin{aligned}\mathcal{L}_{RGM} &= -\sum_{k=1}^{N} \log p(u_k|\mathbf{I}'_k, u_{<k}) \\ &= -\sum_{k=1}^{N}\sum_{t=1}^{T} \log p(u_k^t|\mathbf{I}'_k, u_{<k}, u_k^{<t}).\end{aligned} \tag{2.9}$$

DialoFlow 的总体训练目标为：

$$\mathcal{L} = \mathcal{L}_{CFM} + \mathcal{L}_{SIM} + \mathcal{L}_{RGM}. \tag{2.10}$$

2.3.4 流评分

通过优化上述三个训练目标，DialoFlow 可以捕获对话历史中的动态语义信息流。由于 DialoFlow 模型训练中使用的是人与人的对话，因此可以将语境的变化方向视为对话发展方向的预期。因此，对话机器人对话语句所带来的语义影

响与模型期望之间的距离越近，就意味着越接近人类。基于此，本研究提出了一种基于 DialoFlow 预训练模型的交互式对话自动无参考评估指标——流评分。在人-机器人对话中，当机器人产生一个新的对话语句 u_k 时，流评分测量该对话语句 u_k 所带来的预测语义影响 I'_k 与实际语义影响 I_k 之间的相似性，这可以被认为是该对话语句与人相似的概率。为了计算语义影响之间的相似度，流评分同时测量其余弦相似度和长度相似度:

$$\begin{aligned} s_k &= \cos(\langle \mathbf{I}'_k, \mathbf{I}_k \rangle) \cdot \text{len}(\mathbf{I}'_k, \mathbf{I}_k) \\ &= \frac{\mathbf{I}'_k \cdot \mathbf{I}_k}{||\mathbf{I}'_k||\ ||\mathbf{I}_k||} \cdot \frac{\min(||\mathbf{I}'_k||, ||\mathbf{I}_k||)}{\max(||\mathbf{I}'_k||, ||\mathbf{I}_k||)}. \end{aligned} \tag{2.11}$$

流评分引入长度相似度来考虑模长差异对语义相似度的影响。针对聊天机器人在对话中的整体质量，流评分设计了一个度量，可以看作是对话层面的困惑度:

$$Flow = 2^{-\frac{1}{M}\sum_k^M \log(\frac{s_k+1}{2})}, \tag{2.12}$$

其中 M 表示聊天机器人话语的次数，$\frac{s_k+1}{2}$ 表示相似性值 [0, 1]。流评分（Flow）越低，对话质量越好。

2.4　实验结果与分析

本节将介绍实验采用的数据集、实验设置、基线系统、评估方式、实验结果及实验分析。

2.4.1　实验数据集

对于模型预训练，实验中使用 Reddit 评论，这些评论由第三方收集，并通过 pushshift.io 公开提供 (Baumgartner 等, 2020)。本实验根据 DialoGPT[1] 中使用的处理流程对 Reddit 评论数据进行了清洗。

对于对话回复生成，本实验使用多引用 Reddit 测试数据集 (Zhang 等, 2020b)。该数据集包含 6000 个具有多个参考回复的对话示例。本实验在这个数据集上评估预训练的 DialoFlow 模型。该数据集的对话历史的平均长度为 1.47 轮。为了进一步探讨在长对话历史情况下的动态语义信息流，本实验选择了另一个流行的开放域对话数据集——DailyDialog 数据集 (Li 等, 2017)，该数据集的平均对话历

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

史长度约为 4.66 轮。DialoFlow 预训练模型在 DailyDialog 训练集上进行了微调，并在 DailyDialog 多参考测试集 (Gupta 等, 2019) 上进行了评估。

对于交互式对话自动质量评估，本实验采用了从“第九届国际对话系统技术挑战赛——对话交互式评估赛道”(DSTC9) (Gunasekara 等, 2021) 中收集的数据，其中包含了来自 11 个对话机器人的 2200 次人机对话。对于每个对话，有 3 个人工标注员对整体质量的评分（0-5）。本实验计算了本研究提出的流评分结果与对话机器人级别上的人类评分之间的相关性。由于人与人的对话总是被认为比人与机器人的对话更好，因此本实验也从 BST 数据集 (Smith 等, 2020b) 中随机抽取了 200 个人与人的对话，以查看该指标在真实人类对话中的表现。

2.4.2 实验设置

1. 预训练实验设置

DialoFlow 是基于预训练的 GPT-2 进行预训练的 (Radford 等, 2019)，因为 Zhang 等 (2020b) 表明从预训练的 GPT-2 训练的 DialoGPT 要比从头开始训练的好得多。本研究训练了三种不同的模型尺寸：DialoFlow-base、DialoFlow-medium 和 DialoFlow-large，它们分别是从预训练的 GPT2-base、GPT2-medium、GPT2-large 中训练出来的。本实验使用具有 0.01 权值衰减的 AdamW 优化器 (Loshchilov 和 Hutter, 2019) 和具有 12000 预热步数的线性学习速率调度器。base 和 medium 版本的学习速率为 2e-4，large 版本的学习速度为 1e-4。对于所有模型大小，本实验使用 1024 的批处理大小，对 base 模型和 medium 模型进行最多 4 个 epoch 的训练，对 large 模型进行最多 2 个 epoch 的训练。本实验采用混合精度训练，其中 large 模型在 8 个 Nvidia V100 gpu 上训练大约需要两个月的时间。

2. 对话生成解码设置

在 Reddit 多参考数据集上，本实验对 DialFlow-medium 模型和 DialFlow-large 模型使用波束搜索（波束宽度为 10），在 DialoFlow-base 模型上采用贪婪搜索，与 (Zhang 等, 2020b) 保持相同。在 DailyDialog 数据集上，本实验对预训练的 DialoFlow 和 DialoGPT 进行微调，根据验证集性能选择模型在多参考测试集上进行测试，然后使用波束搜索（波束宽度为 5）进行解码。

表 2.1 对话生成自动评测实验结果

Table 2.1 Automatic evaluation results of dialogue response generation

Method	N-2	N-4	B-2	B-4	METEOR	Entropy	Avg Len
Multi-reference Reddit Dataset							
DialoGPT (B, greedy)	2.39	2.41	10.54%	1.55%	7.53%	**10.77**	12.82
DialoFlow (B, greedy)	2.88	2.93	15.34%	3.97%	9.52%	9.27	**15.43**
DialoGPT (M, beam)	3.40	3.50	**21.76%**	**7.92%**	10.74%	10.48	11.34
DialoFlow (M, beam)	3.89	3.99	20.98%	7.36%	11.46%	10.42	13.37
DialoGPT (L, beam)	2.90	2.98	21.08%	7.57%	10.11%	10.06	10.68
DialoFlow (L, beam)	**3.90**	**4.01**	21.20%	7.42%	**11.48%**	10.42	13.38
Human	3.41	3.50	17.90%	7.48%	10.64%	10.99	13.10
Multi-reference DailyDialog Dataset							
DialoGPT (B, beam)	2.28	2.78	18.83%	6.63%	15.5%	9.80	**18.82**
DialoFlow (B, beam)	3.65	3.84	26.47%	10.12%	16.1%	9.62	12.00
DialoGPT (M, beam)	3.47	3.65	25.39%	9.99%	15.9%	9.64	12.88
DialoFlow (M, beam)	3.80	4.02	27.63%	11.33%	16.7%	9.83	12.06
DialoGPT (L, beam)	3.30	3.46	23.69%	9.20%	15.7%	9.78	13.24
DialoFlow (L, beam)	**3.86**	**4.08**	**28.02%**	**11.57%**	**17.0%**	**9.87**	12.08
Ablation Study on Multi-reference Reddit Dataset							
DialoFlow (M, beam)	3.89	3.99	20.98%	7.36%	11.46%	10.42	13.37
w/o SIM	3.85	3.96	21.36%	7.71%	11.26%	10.43	12.70
w/o SIM & CFM	3.79	3.89	21.33%	7.65%	11.25%	10.33	12.55

2.4.3　基线系统

对于对话回复生成，本实验将本研究提出的 DialoFlow 与 DialoGPT 进行了详细的比较。DialoGPT 是一个在 Reddit 评论中预训练过的流行的对话生成模型。

对于交互式对话评估，本实验将流评分与以下指标进行比较：（1）FED 评分 (Mehri 和 Eskénazi, 2020b)，是一个对话自动质量评估指标，使用 DialoGPT-large，没有任何微调或监督。FED 以 DialoGPT-large 模型为基础，根据几个预先设定的常用人类对话语句，计算出当这些设定对话语句作为后续话语时的困惑度。FED 是在预先设定的人类常用话语下工作的，它可以揭示对话的质量。（2）

表 2.2 对话生成人工评测实验结果

Table 2.2 Human evaluation for dialogue response generation

Metric	DialoFlow	DialoGPT	Tie
Relevance	43.7%	28.8%	27.5%
Informativeness	45.3%	29.2%	25.5%
Human-likeness	46.2%	29.3%	24.5%

表 2.3 流评分对话评估实验结果

Table 2.3 Experimental results for Flow score on dialogue evaluation

Methods	**B1**	**B2**	**B3**	**B4**	**B5**	**B6**
Human ↑	4.142	4.140	4.075	4.035	3.933	3.864
FED ↑	4.988	4.818	4.621	4.670	4.555	4.739
Perplexity ↓	600.0	521.2	441.2	561.6	367.7	1731
Flow ↓	1.396	1.410	1.402	1.406	1.407	1.422
Methods	**B7**	**B8**	**B9**	**B10**	**B11**	**Human**
Human ↑	3.849	3.848	3.828	3.692	3.605	5.000
FED ↑	4.438	4.355	4.651	4.799	3.608	3.468
Perplexity ↓	1879	13347	662.2	618.4	50.29	51.39
Flow ↓	1.425	1.417	1.425	1.461	1.466	1.333

困惑度，用来衡量对话语境下话语的连贯性。本实验使用 DialoGPT-large 模型来测量对话机器人每句话的困惑程度，并把整个对话中所有话语的困惑程度平均做为基准进行度量对话的质量。

2.4.4 评价指标

对于对话回复生成，本实验使用基于通用参考的评价标准：BLEU (Papineni 等, 2002a)、METEOR (Lavie 和 Agarwal, 2007) 和 NIST (Lin 和 Och, 2004) 进行自动评估。NIST 是 BLEU 的一种变体，它通过 n-gram 匹配的信息增益来加权，即间接惩罚无信息的 n-gram，如 “我不知道”，在处理多参考测试集时，这是一个比 BLEU 更适合的度量。本实验还使用熵 (Zhang 等, 2018c) 来评估词汇多样性。本实验使用 DialoGPT 使用的评估脚本进行评估。

对于交互式对话评估，本实验计算自动指标和人工评分之间的皮尔逊相关

表 2.4 流评分、FED 评分、困惑度评分与人工评分相关性

Table 2.4 Correlation between Flow score, FED score, perplexity with human score

Method	Pearson	Spearman
FED	0.67 ($p < 0.1$)	0.56 ($p < 0.1$)
Perplexity	0.12 ($p \approx 0.72$)	0.20 ($p \approx 0.55$)
Flow	**0.91** ($p < 0.001$)	**0.90** ($p < 0.001$)

性和斯皮尔曼相关性，使用预先训练的 DialoFlow-large 模型来计算本研究提出的流评分。

2.4.5　对话生成实验结果

1. 自动评测

表2.1列出了在 Reddit 多参考数据集上预训练的 DialoFlow 与预训练的 DialoGPT 的比较。一般来说，DialoFlow-large 在 NIST 和 METEOR 上得分最高，而 DialoGPT-medium 在 BLEU 上得分更高。DialoFlow 的性能随着模型尺寸的增大而增大，而 Di-aloGPT 在中等尺寸时性能最好。由于 NIST 可以有效地惩罚常见的 n-gram，如 “我不知道”，结果表明，DialoGPT 倾向于生成通用的回复，而 DialoFlow 模型可以生成更多的信息响应。结果还表明，动态流建模有助于提高转换质量，避免生成通用响应。对于词法多样性，DialoFlow 与 DialoGPT 在熵上的表现类似。

多参考 Reddit 数据集的平均历史长度只有 1.45，是较短的。因此，我们在 DailyDialog 数据集（平均对话历史长度 = 4.66）上进行了大量的实验，以验证在长对话历史上的性能增益。如表2.1所示，与 DialoGPT 相比，DialoFlow 显示了在所有模型大小和所有指标上的显著改进。对 DailyDialog 数据集的改进表明，DialoFlow 模型在捕获具有较长历史的动态信息流方面表现出极大的能力。DailyDialog 数据集的性能改进比 Reddit 更显著。本研究认为，Reddit 中的对话主要是论坛中的评论，而 DailyDialog 中的对话则来源于日常生活。因此，在 DailyDialog 数据集中，上下文流在更类似的模式中，与 Reddit 数据集相比，语义影响更容易被预测。

2. 人工评测

本研究使用众包技术对 DailyDialog 测试数据集中随机抽取的 200 个案例进

行人类评估。本研究比较了 DialoFlow-medium 和 DialoGPT-medium。每对回复随机呈现给 3 名人工评估员，他们根据相关性、信息量和类人性对它们进行排序。总体偏好以占总偏好的百分比表示，如表2.4所示，人们对于由 DialoFlow 生成的回复有很强的偏好。人工评估结果表明，对动态信息流进行建模是提高对话生成质量的有效途径。

3. 消融实验

为了探索所提出的训练目标的效果，本研究在 medium 版本的 DialoFlow 上进行了消融实验，如表2.1所示。在这三个训练目标下，DialoFlow 模型在 NIST 和 METEOR 上都取得了最佳的效果。当放弃语义影响建模任务时，性能会略有下降。当进一步放弃上下文流建模任务，这意味着端到端训练，性能再次下降。结果表明，上下文影响建模任务可以有效地进行对话建模，语义影响建模任务可以促进上下文影响建模任务。

2.4.6 对话评估实验结果

表2.4显示了国际对话技术挑战赛 DSTC9 交互式对话数据集上不同自动评价指标与人类评分的聊天机器人级别的相关性。本研究提出的流评分与人工评分达到了很强的斯皮尔曼相关性：0.90 ($p<0.001$) 和很强的皮尔逊相关性：0.91 ($p<0.001$)。FED 评分与人工评分仅表现为中度相关性，聊天机器人水平的斯皮尔曼相关性为 0.56 ($p<0.1$)。困惑度得分表现出非常弱的相关性。一方面，结果表明，本研究提出的流评分可以有效地估计聊天机器人的整体质量。另一方面，高相关性也表明，基于预知未来的 DialoFlow 模型捕获了人与人自然对话中的动态信息流。

表2.3显示了在 DSTC9 交互式对话评估跟踪中 11 个聊天机器人的详细的人类评分、FED 评分、困惑度和流评分。好的自动指标不仅应该在人与聊天机器人的对话中表现良好，也应该在人与人的对话中表现良好，因为聊天机器人的最终目标是生成类似于人类的响应。与其他 11 个聊天机器人相比，FED 在人机对话上的表现很差。本研究提出的流评分是将人与人之间的对话作为最好的一种，而人与最好的聊天机器人之间的流评分差距与人的评分差距类似。

流评分可以看作是话语层面上的困惑。在自然会话中，针对某一特定语义有许多不同的表达方式。传统的词级困惑度可以判断话语的连贯和流畅性，但在多样性表达上表现不稳定。流评分直接度量语义相似度，解决了传统困惑问题。

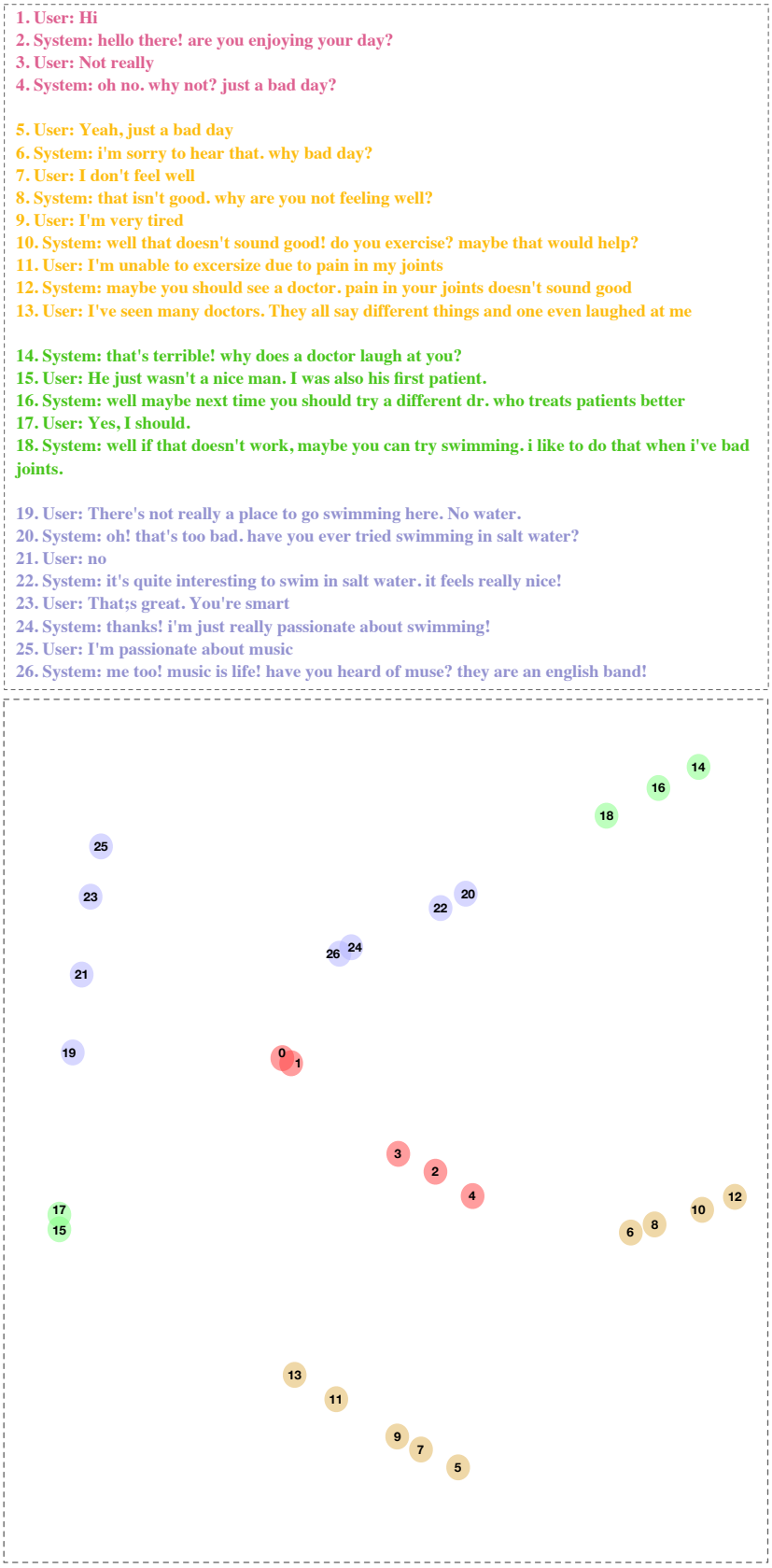

图 2.3 一个人机对话案例的高层语义信息流的 2D T-SNE 可视化

Figure 2.3 2D T-SNE visualization of the semantic context of a human-bot conversation

2.4.7 案例研究

图2.3显示了由预训练的 Dialoflow 模型编码的人机对话的高层语义信息流的 2d T-SNE 可视化。由图2.3可见，对话可以分成四个主题：问候（1~4），谈论为什么糟糕的一天（5~13），解释看医生的糟糕经历（14~18），以及讨论游泳(19~26)。与此相对应的是，在可视化中，当话题切换时，高层语义信息流发生了很大的变化，这表明 Dialoflow 能够捕捉对话中的高层语义动态信息流，并有效地衡量每个对话语句所带来的语义影响。此外，不同的说话者可以保持着他们自己独特的语义流。

2.5 本章小结

本研究针对对话逻辑混乱问题，引入高层语义信息流建模对话历史，提出了 DialoFlow 预训练对话模型，通过处理每个对话语句所带来的语义影响来模拟对话语句之间的动态信息流。具体来说，DialoFlow 模型采用了一个单向流模块来建模上下文的语义变化，并通过三个训练目标来进行优化。此外，在 DialoFlow 的基础上，本研究提出了流评分，一种基于预训练的 DialoFlow 的交互式对话评价的自动无参考评估指标。在对话回复生成上的实验表明，该方法能够有效地捕捉话语间的动态信息流，提升对话历史建模的效果，提高对话的前后文连贯性。在对话评估上的实验表明，流评分可以较好地对开放域交互式对话进行前后逻辑性上的评分。在未来的工作中，本研究希望将 DialoFlow 应用于面向任务的对话，并探索在长文本生成（如故事生成）上的应用。

第 3 章　引入多模态知识的开放域对话回复生成研究

3.1　引言

人们在对话过程中会从周围的环境中获取多模态的信息，例如视觉信息、听觉信息、文本信息等。这些信息帮助人们进行接下来的对话。相关信息的缺失会限制人们对对话内容的理解以及回复正确的对话内容。同样，对于开放域聊天机器人，缺乏相关信息，会生成较多的通用回复或产生幻觉回复，与真实场景相悖。为了使开放域对话机器人具备理解多模态信息的能力，生成流畅、有信息量的回复，本章进行了引入多模态知识的开放域对话研究。特别地，本章研究了基于视听场景感知的开放域对话生成，引入视频、音频、文本知识辅助开放域对话回复生成。

近年来，视听场景感知开放域对话生成因其广泛的应用而越来越受到工业界和学术界的关注。Zhou 等 (2018b) 提出了一个基于电影文档的文本对话数据集。Urbanek 等 (2019) 搭建了一个大型的文本冒险游戏平台，在这个平台上，对话机器人可以根据文本中描述的场景进行行动和对话。受人类固有的多模态知识获取和理解能力的启发，AlAmri 等 (2018) 将多模态信息整合到场景感知对话中，提出了视听场景感知开放域对话（Audio-Visual scene-aware Dialog, AVSD）任务。这些工作的目的是在给定场景的基础上产生信息丰富和流畅的对话回应。视听场景感知对话任务的目标是通过理解所有形式的信息（如文本、视频和音频）来生成正确流畅的回答，这比基于图像或基于纯文本的对话任务更具挑战性。图3.1展示了视听场景感知开放域对话数据集 (AlAmri 等, 2018) 中的示例对话。该任务面临两个挑战：(1) 获取准确的多模态表示，并在不同模态之间进行有效交互；(2) 更好地利用多模态信息辅助理解对话并产生对话回复。

对于第一个挑战，最近的一些工作对如何获取准确的多模态表示进行了大量探索。为了获得准确的多模态表示，现有的方法首先利用独立编码器对不同的模态进行单独编码，然后利用注意机制融合不同模态的表示。但是，单独编码单模态的信息不能从其他模态的相关信息中获益。例如，视听场景感知中的文本信息，如视频字幕，对理解视频是有帮助的。

对于第二个挑战，场景感知的对话回复生成，也是相当困难的。对话模型需

Caption: A woman standing in a hallway takes off her slippers. She then climbs on a chair and starts doing something with the ceiling light.

Summary: A woman about 30 years old wearing a jean skirt and top is standing on a stool and fixing something in the hallway next to a door. The hallway has linoleum floors.

Q1: where is the video happening ?

A1: it is happening inside in the hallway

Q2: are there any people in the video ?

A2: yes there is one person in the video.

…

Q10: what is the person doing ?

A10: she is standing on a stool doing something with the ceiling light.

图 3.1 视听场景感知的开放域对话示例

Figure 3.1 An example of audio-visual scene-aware open-domain dialogue.

要完全理解给定场景的对话历史，并捕获对话回合之间的相关依赖关系，以生成信息丰富且正确的对话回复。此外，构建大规模场景感知对话数据集的成本较高，且仅在该任务的数据集上训练生成模型，性能有限。采用预训练的语言模型可以利用从其他文本数据中学习到的丰富的语言依赖关系来改进规模有限的场景感知对话数据集。

为了解决上述问题，本文设计了一种通用的多模态 Transformer 模型，将不同的模态联合编码并同时产生对话回复。受 Bert (Sun 等, 2019)、GPT-2 (Radford 等, 2019) 等预训练工作的启发，本文采用自监督学习方法，并采用多任务学习 (对话回复语言建模、视频音频序列建模和字幕语言建模) 方法学习联合表示，并生成信息丰富、流畅的对话回复。由于通过利用大规模的预训练的语言模型，在许多下游对话生成任务中取得了巨大的成功，本研究扩展了预先训练的 GPT-2 (Radford 等, 2019) 模型，通过将视觉和文本表示结合到一个结构化序列中，并对其进行微调，以捕获跨模态依赖并生成对话回复，来应对上述挑战。

3.2　相关工作

3.2.1　知识驱动的开放域对话系统

大多数关于对话系统的工作集中在开放领域对话或面向任务的对话。就像人与人之间的对话一样，总是会有背景知识。最近的一些研究工作可以根据文档或结构化知识图 (Li 等, 2019b; Zhou 等, 2018b; Reddy 等, 2019; Dinan 等, 2019; Madotto 等, 2018) 生成对话回复。这些系统可以产生与背景知识更相关或更正确的对话回复。有一些研究工作将多模态信息融入问答和对话中。在视觉问答任务 (Goyal 等, 2017; Agrawal 等, 2017) 中，系统的目标是回答关于图像内容的给定问题。视觉对话 (Das 等, 2017) 的任务是根据给定的图像和对话上下文在对话中生成自然的对话回复。这些研究工作将文本或图像作为背景知识，而在视听场景感知对话中，知识是文本、视频和音频。Hori 等 (2019) 引入了一种基于长短期记忆网络的多模态编码器和解码器。Nguyen 等 (2018) 提出了一种基于视频的视听特征提取器的递阶循环编解码框架。Pasunuru 和 Bansal (2019) 采用双重注意力机制对多种模式进行编码和对齐。DSTC7-AVSD 比赛的获胜团队 (Sanabria 等, 2019) 使用了层次注意力机制结合文本和视觉模态，并使用 How2 数据集对视频编码器进行预训练。此外，MTN (Le 等, 2019) 提出了多模态 Transformer 模型来对视频进行编码，并采用来自不同模态的信息。

3.2.2　基于大规模预训练语言模型的开放域对话系统

研究表明，大规模预训练语言模型在提高语言生成任务（如对话系统和文本摘要）的性能方面发挥着重要作用。Zhang 等 (2019a) 提出了一种基于 BERT 的自然语言生成模型，在编码和解码过程中充分利用了预训练的语言模型。Wolf 等 (2019b) 使用生成式预训练 Transformer 模型将迁移学习引入生成式数据驱动对话系统。本研究将这种迁移学习方法扩展到引入多模态信息的对话生成任务中，并提出了一种自监督学习方法来更好地进行多模态表示与融合。

3.3　基于预训练语言模型的通用多模态 Transformer 模型

视听场景感知开放域对话任务的目标是基于多模态知识（包括视频、音频、视频描述和对话上下文）的信息生成流畅、有信息量的对话回复。形式上，设 $\mathbf{V}$ 和 $\mathbf{A}$ 分别表示视频和音频。考虑到对话摘要和视频描述之间的相似性，本研究

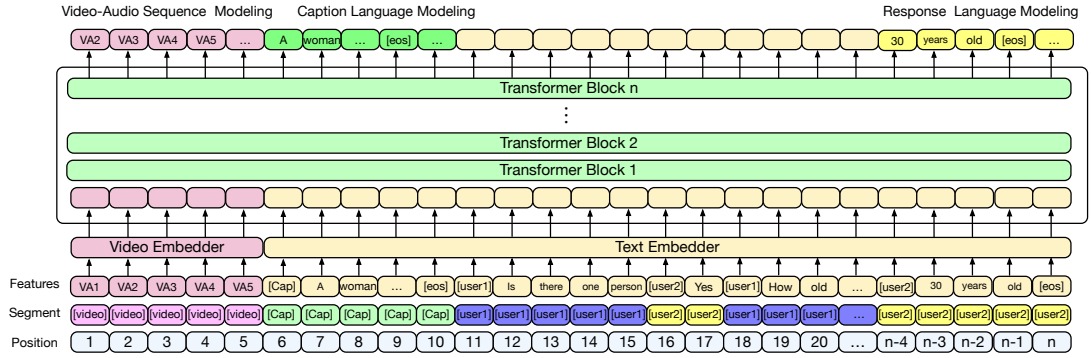

图 3.2 基于预训练语言模型的通用多模态 Transformer 模型

Figure 3.2 Universal multi-modal transformer based on the pre-trained language model

将对话摘要和视频描述连接成一个完整的视频描述 $\mathbf{C} = \{c_1, c_2, \ldots, c_I\}$，它通常提供视频和整个对话的总结。使用 $\mathbf{U} = \{\mathbf{Q}_1, \mathbf{R}_1, \mathbf{Q}_2, \mathbf{R}_2, \ldots \mathbf{Q}_N, \mathbf{R}_N, \}$ 表示对话的 N 个回合，其中 $\mathbf{Q}_n$ 表示问题 N，$\mathbf{R}_n = \{r_n^1, r_n^2, \ldots, r_n^m\}$ 表示包含 m 个单词的对话回复 $\mathbf{R}_n$。因此，考虑视频 $\mathbf{V}$、音频 $\mathbf{A}$、对话历史 $\mathbf{U}_{<n}$ 和视频描述 $\mathbf{C}$，对于给定问题 $\mathbf{Q}_n$ 生成对话回复 $\mathbf{R}_n$ 的概率可计算为:

$$P(\mathbf{R}_n|\mathbf{V}, \mathbf{A}, \mathbf{C}, \mathbf{U}_{<n}, \mathbf{Q}_n; \theta) = \prod_{j=1}^{m} p(r_n^j|\mathbf{V}, \mathbf{A}, \mathbf{C}, \mathbf{U}_{<n}, \mathbf{Q}_n, r_n^{<j}; \theta) \tag{3.1}$$

本研究提出了一个统一的多模态 Transformer 模型，模型结构如图3.2所示。整体上讲，该模型是一个基于 GPT-2 模型结构 (Radford 等, 2019) 的多层 Transformer 模型。针对视听场景感知对话数据资源稀缺问题与多模态表示融合问题，本研究通过 GPT-2 初始化一个 12 层具有多头自注意力机制的的 Transformer 解码器模型，使用视频音频嵌入将视频音频特征映射到文本空间，并设计了三种任务来促进多模态融合：对话回复生成建模、视频音频序列建模、视频描述生成建模。

3.3.1 模型输入

1. 文本输入

对于文本特征，遵循 GPT-2 (Radford 等, 2019) 的方式，将输入的句子使用 WordPieces (Wu 等, 2016) 工具进行分词。

2. 视频与音频输入

对于给定的视频 V_k，本研究使用 l 帧的滑动窗口将视频分割为 T_k 段。如图3.3所示，对于每个分段 $S_t = \{f_1, f_2, \ldots, f_l\}$，其中 f_i 表示一帧，使用一个预

训练的 I3D-rgb 模型和 I3D-flow 模型 (Kay 等, 2017a) 提取 d_v 维视频特征 $\mathbf{V}_{rgb}$ 和 $\mathbf{V}_{flow}$。考虑到音频与视频是同步的，我们从同一段视频中选取对应的音频，使用预先训练的 VGGish 模型 (Hershey 等, 2017) 提取 d_a-维音频特征 $\mathbf{A}_{vggish}$。然后拼接视频 I3D-rgb 特征、I3D-flow 特征和 VGGish 特征:

$$\mathbf{V}_t = [\mathbf{V}_{rgb}, \mathbf{V}_{flow}, \mathbf{A}_{vggish}], \mathbf{V}_t \in \mathbb{R}^{2d_v+d_a} \tag{3.2}$$

然后本研究将视频音频特征 $\mathbf{V}$ 送入如图 3.3所示的视频编码全连接层 (Video Embedder)，投影到与文本嵌入相同的嵌入空间。如图3.3所示，本研究提出的模型根据不同的段嵌入 (SE) 区分不同部分的输入 (视频输入、视频描述输入、说话人 1 对话语句和说话人 2 对话语句)，最终每个词的表示由其词嵌入 (WE)，位置编码嵌入 (PE) 和段嵌入 (SE)。注意，“[video]”，“[cap]”，“[user1]” 和 “[user2]” 分别用来表示视频音频片段、视频描述、说话人 1 对话语句和说话人 2 对话语句。

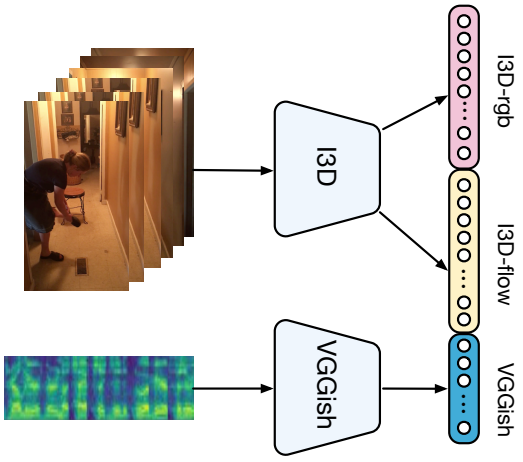

图 3.3 视频与音频特征提取

Figure 3.3 Feature extraction for video and audio

3.3.2　多任务学习

本研究设计了三个任务来训练模型：以视频、音频、视频描述和对话历史为条件的对话回复生成建模、以视频描述和对话为条件的视频-音频序列建模和以视频和音频为条件的视频描述生成建模。

1. 对话回复生成建模

这个任务的目标是生成对话回复 $\mathbf{R}_n = \{r_n^1, r_n^2, \ldots, r_n^m\}$，基于视频音频特征 $\mathbf{V}$，

视频描述 $\mathbf{C}$，对话历史 $\mathbf{U}_{<n}$，对话问题 $\mathbf{Q}_n$，最小化负对数似然损失函数：

$$\mathcal{L}_{RLM}(\theta) = -E_{(\mathbf{V},\mathbf{C},\mathbf{U},\mathbf{Q},\mathbf{R})\sim D} \log \prod_{j=0}^{m} p(r_n^j | \mathbf{V}, \mathbf{C}, \mathbf{U}_{<n}, \mathbf{Q}_n, r_n^{<j}) \tag{3.3}$$

其中，$r_n^{<j}$ 表示对话回复 $\mathbf{R}_n$ 的前 j 个词，θ 表示可训练参数，$(\mathbf{V}, \mathbf{C}, \mathbf{U}, \mathbf{Q})$ 是从整个训练集 D 中采样得到的。

2. 视频-音频序列建模

这个任务是在给定视频描述和对话历史时预测视频音频特征。与以离散标签表示的文本标记不同，视频音频特征是高维的和连续的。区别于Sun 等 (2019) 将视频音频特征聚类成离散标签，本研究采用了 (Chen 等, 2020) 的视频音频特征回归方法。本任务将 Transformer 模型的视频音频表示输出还原到下一个视频音频特征输入 $\mathbf{V}_{t+1}$。具体地，本研究应用一个全连接层来将输出转换为与视频音频输入 $\mathbf{V}_{t+1}$ 相同维度的向量 $g_\theta(\mathbf{o}_t)$。通过最小化 L2 损失来训练这个任务：

$$\mathcal{L}_{VASM}(\theta) = E_{(\mathbf{V},\mathbf{C},\mathbf{U})\sim D} \frac{1}{T} \sum_{t=1}^{T} \| g_\theta(\mathbf{o}_t) - \mathbf{V}_{t+1} \|_2^2 \tag{3.4}$$

其中，$\mathbf{o}_t = f_\theta(\mathbf{V}_{<t+1}, \mathbf{C}, \mathbf{U})$，$f_\theta$ 代表 GPT-2 模型的作用。

3. 视频描述生成建模

与对话回复生成建模任务类似，基于视频音频特征 $\mathbf{V}$ 通过最小化负对数似然损失函数训练模型生成 $\mathbf{C} = \{c_1, c_2, \ldots, c_I\}$：

$$\mathcal{L}_{CLM}(\theta) = -E_{(\mathbf{V},\mathbf{C})\sim D} \log \prod_{i=0}^{I} P(c_i | \mathbf{V}, c_{<i}) \tag{3.5}$$

3.4 实验结果与分析

3.4.1 数据集

本研究使用来自国际对话技术比赛 DSTC7 和 DSTC8 的视听场景感知对话 (AVSD) 数据集 (Hori 等, 2019)。在这个数据集中，每个对话有两个参与者，一个提问者和一个回答者。每个对话由一系列关于给定视频的问题和答案组成。每个视频都有一个视频描述和一个对话摘要。视频描述是对给定视频的大致描述。对话摘要是对基于此视频的对话的总结。本实验使用了最先进的视频特征提取器 I3D 模型和 Kinetics 数据集 (Kay 等, 2017b) 进行视频特征提取。具体来说，实

验中使用 I3D 网络的 “Mixed 5c” 层的输出（一个 2048 维的向量）作为视频的表示。对于音频特征，实验采用了著名的 VGGish 模型 (Hershey 等, 2017)，该模型输出 128 维嵌入。视听场景感知对话数据集有 7659 个训练对话，1787 个验证对话，1710 个测试对话。在 DSTC7 和 DSTC8 中数据集中，训练集和验证集是相同的，而测试集是不同的。在两个测试数据集上对本研究提出的方法进行评估。

3.4.2　基线模型

本研究将提出的模型和方法与几种相关的基线方法进行比较：数据集基线模型，多层次注意力模型，MTN 系统等:

- 数据集基线系统：由数据集提出者提供的多模态基线，将所有模态与投影矩阵结合在一起 (Hori 等, 2019)。
- 多层次注意力模型：层次注意力方法结合文本和视觉模式。这是 DSTC7-AVSD 任务中排名第一的团队使用的方法。
- MTN (Le 等, 2019)：在 DSTC8-AVSD 挑战之前的最先进的系统，提出了多模态 Transformer 网络 (MTN) 来编码视频和引入来自不同模态的信息。
- JMAN (Chu 等, 2020)：基于 RNN 的多步多模态融合注意力网络，实现了多步注意力机制，同时考虑了视觉和文本的表示。
- MSTN (Lee 等, 2020)：MSTN 采用基于 Transformer 的模型结构，并在词的生成阶段考虑重点词的含义，采用基于注意力机制的词嵌入层。
- STSGR (Geng 等, 2020)：本工作将一个视频表示为两个时空场景图，通过图注意力编码，并使用多模态 Transformer 进行高层信息推理。

3.4.3　评估指标

1. 客观评价

实验使用了在自然语言生成任务中常用的度量，如 BLEU (Papineni 等, 2002b)、METEOR (Denkowski 和 Lavie, 2014)、ROUGE-L (Lin, 2004) 和 CIDEr (Vedantam 等, 2015)。这些评价指标被用来计算预测的对话回复和标准对话回复之间的词重叠程度。实验使用了 DSTC8 视听场景感知开放域对话生成比赛组织者提供的工具包来进行评估。

2. 主观评价

主观评价是对话回复生成评测的必要条件。组织者对一些系统进行了评估，

这些系统是基于众包的人工评分。评注者被要求考虑所生成的回答的正确性、自然性、信息量和合适性，并给五个等级打分，从 0 到 4。人类生成的标准参考对话回复的评分是 4.000。

3.4.4 实验设置

实验使用 GPT2-base 模型 (Radford 等, 2019; Wolf 等, 2019a) 的预训练的权值来初始化模型。训练过程使用了最多 3 轮的对话历史。模型的隐藏尺寸为 768，训练批大小为 32，Adam 优化器学习速度为 6.25e-5。解码过程使用波束大小为 5 的波束搜索，最大回复长度为 20，长度损失为 0.3。对于视听场景感知开放域对话任务，有三种不同的设置：纯文本、文本 + 视频和不包含视频描述的文本 + 视频。纯文本设置等同于基于文本知识的对话生成任务。文本 + 视频设置是一个完整的基于文本和视觉听觉信息的场景感知对话生成任务。不包含视频描述的文本 + 视频主要集中在只有视频但没有字幕的情况下，这是现实世界中的典型任务。

3.4.5 实验结果

本节展示了三种不同设置下的实验结果：纯文本、文本 + 视频和不包含视频描述的文本 + 视频。

1. 纯文本

纯文本设置只使用文本输入，其中包括对话历史、视频描述。该实验中只使用对话回复生成建模（RLM）任务来训练模型。这些结果在表3.1和表3.2的 “text only” 设置的 “Our model（RLM）” 行中展示。如表3.1所示，与分层注意力（在 DSTC7-AVSD 挑战的获胜系统）和 JMAN 相比，本研究提出的模型在 DSTC7-AVSD 数据集上的所有指标上都获得了更好的性能。其中，本研究提出的模型的 BLEU-4 提高了 0.069，CIDEr 提高了 0.185。此外，表3.2还显示了 DSTC8-AVSD 测试数据集的人工评价结果。在人的评价过程中，评价者被要求对模型的回复以及标准回复进行打分，其中标准回复的得分为 4.000。本研究的模型得分为 3.934，这是所有 DSTC8 视听场景感知的开放域对话生成比赛提交的结果中最高的人类评分。从人类评价的角度来看，本研究提出的模型结果与人类对话非常接近。

2. 文本 + 视频

在该设置中，本研究使用文本输入 (对话历史及视频描述) 和视频音频输入，

表 3.1 在 DSTC7-AVSD 测试集上的客观评价实验结果

Table 3.1 Objective evaluation results on the DSTC7-AVSD test set

Models	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
Input: text only							
Hierarchical Attention	-	-	-	0.376	0.264	0.554	1.076
Our model (RLM)	**0.747**	**0.627**	**0.527**	**0.445**	**0.287**	**0.594**	**1.261**
Input: text + video							
Hierarchical Attention	-	-	-	0.394	0.267	0.563	1.094
MTN	-	-	-	0.392	0.269	0.559	1.066
Our model (RLM)	0.759	0.635	0.533	0.448	0.293	0.602	1.282
+ VASM	**0.765**	**0.643**	**0.543**	**0.459**	**0.294**	**0.606**	**1.308**
Input: text + video w/o caption / summary							
Naive fusion	-	-	-	0.309	0.215	0.487	0.746
DSTC7-AVSD Team 9	-	-	-	0.315	0.239	0.481	0.773
Our model (RLM)	**0.694**	**0.570**	**0.476**	**0.402**	**0.254**	**0.544**	**1.052**
+ VASM	0.677	0.556	0.462	0.389	0.250	0.533	1.004
+ recaption	0.670	0.537	0.438	0.362	**0.254**	0.535	1.022

并使用对话回复生成建模（RLM）任务和视频音频序列建模（VASM）任务训练模型。这些结果在表3.1、表3.2“text+video” 设置的 “Our model (RLM) + VASM” 行中展示。从表3.1中可以看出，与 MTN（视听场景感知开放域对话生成任务之前的最先进的模型）相比，本研究提出的模型也有了很大的改进。具体地，本研究提出的模型将 BLEU-4 分数提高了 0.056，将 CIDEr 分数提高了 0.216。与纯文本任务相比，本研究提出的模型在客观评价上取得了更好的效果，这表明该视频理解方法是有效的。如前所述，本研究采用多任务学习 VASM 任务使 BLEU-4 评分提高了 0.011 分，CIDEr 评分提高了 0.026 分。在表3.2的 DSTC8 测试数据集结果中，该方法比 CIDEr 评分提高了 0.014，说明该方法是有效的。通过大量的案例研究，“text+video” 的人类表现在理论上优于 “text only”。

3. 不包含视频描述的文本 + 视频

在这个设置中，有两种方法：（1）在训练和测试中，不使用视频描述。用对话回复生成建模 (RLM) 和视频音频序列建模 (VASM) 任务对模型进行训练。结果在表3.1、表3.2的 “Our model (RLM) + VASM” 设置中 “text + video w/o caption

表 3.2 DSTC8-AVSD 测试集上客观评价与主观评价实验结果

Table 3.2 Objective and subjective evaluation results on the DSTC8-AVSD test set.

Models	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	Human rating
Input: text only								
Our model (RLM)	0.744	0.626	0.525	0.442	**0.287**	0.595	1.231	**3.934**
Input: text + video								
Our model (RLM)	0.739	0.624	0.528	**0.447**	0.284	0.592	1.226	3.895
+ VASM	**0.746**	**0.626**	**0.528**	0.445	0.286	**0.598**	**1.240**	-
Input: text + video w/o caption / summary								
Our model (RLM)	**0.677**	**0.556**	**0.462**	**0.387**	**0.249**	**0.544**	**1.022**	-
+ VASM	0.669	0.550	0.457	0.385	0.246	0.540	0.988	-
+ recaption	0.661	0.533	0.437	0.364	0.242	0.533	0.991	-

and summary" 进行展示。(2) 在训练中，使用视频描述，并按照之前模型中描述的三个任务对模型进行训练。测试时，首先根据给定的视频音频输入生成视频描述，然后使用视频音频输入、生成的视频描述和对话历史生成对话回复。结果展示在表3.1、表3.2设置的 "Our model (RLM) + VASM + CLM" 中。

这种设置是最类似于现实世界的视听场景感知开放域对话：只有视频音频信息和对话历史。因此，这项任务更具挑战性。如表3.1所示，我们发现文本 + 视频任务的性能低于预期，但令人欣慰的是，本研究提出的模型仍然表现得比较好，表现远远超过基线模型。在这个任务中，实验也尝试使用多任务学习方法，包括视频音频序列建模（VASM）和视频描述生成建模 (CLM)，但这导致了几乎所有指标的性能降低。这些现象将在下一节进行分析讨论。

3.4.6 实验分析

1. *训练方法分析*

正如实验结果所展示的，在采用视频音频序列建模后，本研究提出的模型在文本 + 视频任务中性能较好，但在文本 + 视频无视频描述任务中性能较差。本研究分析认为原因在于，如果只使用对话历史而没有视频描述，很难重建被遮挡的视频的特征。与从相邻的文本中重建文本相比，本研究认为该模型在从视频中

表 3.3 对话历史长度对模型性能的影响

Table 3.3 Influence of different dialogue history lengths

History Length	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
0	0.729	0.599	0.496	0.413	0.275	0.573	1.182
1	0.760	0.638	0.536	0.452	**0.296**	0.605	1.305
2	0.755	0.632	0.532	0.450	**0.296**	0.601	1.297
3	**0.765**	**0.643**	**0.543**	**0.459**	0.294	**0.606**	**1.308**
5	0.758	0.634	0.533	0.451	0.292	0.601	1.293
9	0.759	0.631	0.526	0.441	**0.296**	0.603	1.294

表 3.4 不同解码方式在客观评价上的比较结果

Table 3.4 Comparison of the performance with different decoding methods on the objective evaluation

Decoding Methods	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
Greedy Search	0.743	0.610	0.503	0.416	0.284	0.587	1.217
Nucleus Sampling	0.680	0.525	0.410	0.321	0.252	0.527	0.955
Beam Search	**0.765**	**0.643**	**0.543**	**0.459**	**0.294**	**0.606**	**1.308**

提取信息方面并没有很强的能力。因此，该方法在不包含视频描述的文本 + 视频设置中表现得不是很好。对于视频描述生成任务（CLM）性能较差的原因，本研究认为可能是相似的：模型提取视频特征的能力有限，限制了从视频中推断视频描述的性能。因此，本研究认为未来的工作可以更多地关注于视频音频理解，以获得更好的基本视频音频特征。

2. 对话历史长度分析

本研究使用模型在带有视频-音频序列建模（VASM）任务的文本 + 视频设置中进行实验，以探讨对话历史长度的影响。如表3.3所示，当最大对话历史长度为 3 时，模型通常表现最好。

表 3.5 视听场景感知开放域对话生成的案例

Table 3.5 Case study of dialogue response generation

视频描述	A man with glasses wearing a green t-shirt is playing with a device on a counter top in a kitchen.
对话历史	[User1]: So, this person is standing in that spot for the entire length of time, fiddling with something? [User2]: He stands there and flips the item around, and taps it, then at the very end, he goes and closes the fridge door. [User1]: Is this person wearing glasses?
基线模型	[User2]: No, they are not wearing glasses.
纯文本模型	[User2]: Yes, he is wearing glasses.
文本 + 视频模型	[User2]: Yes, he is wearing glasses.
参考回复	[User2]: Yes, he is wearing glasses.
视频描述	A person stands in a bedroom holding a folded sheet. They are at a panel of switches. They turn on the light.
对话历史	[User1]: Does the person hold it during the whole video? [User2]: Yes, while standing at the panel of switches, he is holding the folded sheet the entire time. [User1]: What does the man touch with his left hand on the wall?
基线模型	[User2]: He doesn't touch anything.
纯文本模型	[User2]: He touches with his right hand on the wall.
文本 + 视频模型	[User2]: He flips a light switch on.
参考回复	[User2]: He is touching a panel of switches and he turns on the light.
视频描述	A man is laying in bed watching tv. He mumbles something and then gets up and begins vacuuming the room.
对话历史	[User1]: In the door way to the next room, there is an object. What is that ? [User2]: He went to the doorway for a vacuum. [User1]: When he gets up, does he have anything in his hands?
基线模型	[User2]: No he is not talking.
纯文本模型	[User2]: He has a vacuum in his hand.
文本 + 视频模型	[User2]: No, he has nothing in his hands.
参考回复	[User2]: He had nothing in his hands when he got up.

3. *解码方法分析*

为了找到一种有效的视听场景感知开放域对话生成的解码方法，实验尝试了多种译码方法，包括贪心搜索、波束搜索和原子采样，该方法从概率分布的动态核中采样文本，通常用于生成多种文本。如图3.4所示，在三种译码方法中，波束搜索译码在所有自动评测指标上的效果都最好。我们认为，视听场景感知开放域对话，其回复相对于无场景限制的开放域对话更为明确。因此，在该任务中解码时最好采用波束搜索。

3.4.7　案例分析

表3.5比较了由基线模型、模型（纯文本）和模型（文本 + 视频）生成的对话回复。与基线模型相比，本研究提出的模型可以生成具有更多的信息量的对话回复。如案例 1 所示，对于纯文本模型，当可以在视频描述中找到信息时，对话回复可以生成的得很好。然而，如案例 2 和案例 3 所示，当涉及到询问视频描述中没有出现的特定信息时，纯文本模型表现得并不好。在这些情况下，文本 + 视频模型可以根据视频音频信息，查找相关信息，并生成正确的对话回复。

3.5　本章小结

针对开放域对话系统缺乏相关背景知识问题，本研究在开放域对话对话中引入了多模态知识，进行了试听场景感知的开放域对话对话生成研究。为了缓解多模态对话资源稀缺，多模态表示融合难问题，本研究提出了一种基于预训练语言模型的通用的多模态对话生成模型，并设计了三个任务来训练模型：对话回复生成建模，视频音频序列建模，视频描述生成建模。通过这些任务，模型可以在多种模态中学习更准确的融合表示，并生成具有更多信息量的对话回复。本研究的系统在 DSTC7-AVSD 和 DSTC8-AVSD 测试数据集的客观评价中都取得了最佳的性能。在 DSTC8-AVSD 测试数据集中，本研究的系统生成的对话回复的人工评估的评分相比于标准回复的人工评分高达 98.4%。在未来，我们计划使用更多的视频特征，如 ResNet 特征，并探索更多的训练任务，以提高视频、音频和文本的联合表示的效果。此外，我们希望将这些方法扩展到其他任务，如视频描述生成、图像描述生成和更通用的场景感知开放域对话。

第 4 章　引入问询的开放域对话前后一致性自动评估研究

4.1　引言

近些年来，随着大规模预训练技术的发展 (Devlin 等, 2019b; Radford 等, 2019; Brown 等, 2020b) 与大规模高质量对话数据的构建 (Dinan 等, 2019; Baumgartner 等, 2020; Smith 等, 2020b)，开放域智能聊天机器人 (Adiwardana 等, 2020a; Bao 等, 2020a,b; Smith 等, 2020a; Li 等, 2021e) 取得了长足的进步。这些成果无疑是非常令人激动的，但是构建一个真正像人的开放域聊天机器人还有很长的路要走。

DialoGPT 762M	
Human:	*What movies do you like most?*
DialoGPT:	*The Unforgiven from Metallica.*
Human:	*What do you think of the Unforgiven?*
DialoGPT:	*I never heard of it.*
Blender 2.7B	
Human:	*What do you like to cook?*
Blender:	*I only cook Indian cuisine. How about you?*
Human:	*I enjoy cooking Chinese food, especially the dumplings.*
Blender:	*Chinese food is delicious. I also like cooking the Chinese food.*
Plato 1.6B	
Human:	*Do you like dogs?*
Plato:	*Yes, dogs are cool, I prefer cats though.*
Human:	*Do you have a dog?*
Plato:	*I do not, but I'd love a dog. I think dogs are so cute.*
Human:	*I love dogs very much. I have three dogs.*
Plato:	*I have three dogs too. One puppy and two Huskies.*

图 4.1 目前常用聊天机器人在人机交互过程中出现前后不一致的现象

Figure 4.1 Several human-bot conversations demonstrate that popular chatbots generate inconsistent responses when talking to a human under some specific conditions

目前，通过建模多轮对话逻辑与引入相关背景知识，开放域聊天机器人在交互过程中可以较好地生成流畅的、上下文相关的、吸引人的、有知识性的回复，但是在对话前后一致性上还存在很多不足之处 (Nie 等, 2020a)。对话前后一致性是指聊天机器人产生的对话回复不能与对话历史发生冲突。图4.1展示了一些常

用的开放域聊天机器人（DialoGPT (Zhang 等, 2020b)，Blender (Smith 等, 2020b)，Plato (Bao 等, 2020a)）在人机交互过程中出现前后不一致的例子。根据本研究中对聊天机器人人机交互过程的观察，很多开放域聊天机器人在与人交互的过程中很容易与对话历史产生冲突，本文称之为对话前后不一致。对话前后不一致会很大程度上影响人机交互的体验，降低交互的满意度。因此，保证对话前后一致性是开放域聊天机器人交互的基本要求。然而，目前开放域对话前后一致性提高是比较困难的，重要原因是因为缺乏一种有效、高效、稳定的对话前后一致性评估方法。

为了评估开放域聊天机器人的前后一致性，最直接的方法是依靠人工标注员判断聊天机器人生成的对话是否具有前后一致性。但是，人工标注员经常是临时雇佣的，并且评测标准与人类主观较为相关，无法清晰定义。这就导致了在进行人工评测过程中，人工标注员之间经常得出相互不一致的结果 (Mehri 和 Eskénazi, 2020c)。为了解决这个问题，一些研究工作提出开发自动评测方法 (Welleck 等, 2019b; Song 等, 2020b; Nie 等, 2020b)。这些方法在检测对话中的前后冲突问题上是有效的，但是这些方法是基于人机对话数据。人机对话数据的采集过程是非常耗时耗力的，并且容易产生很多低质量的对话数据 (Deriu 等, 2020; Dinan 等, 2020b)。所有这些问题严重阻碍了对话前后一致性评估的进行以及聊天机器人前后一致性的提升。因此，为了提升效率、降低人工主观偏见，本研究采用聊天机器人与聊天机器人进行对话交互的方式产生对话数据。

当足够的聊天数据产生后，现有工作大多采用 Nie 等 (2020b) 的方法进行对话前后不一致检测，即检测当前对话回复与所有对话历史是否发生冲突。然而，对话前后不一致现象的出现次数相比总体对话占比较低，检测当前对话回复与所有对话历史是否发生冲突引入了大量噪声，并且检测效率较低。基于对大量的聊天机器人对话交互过程的观察，本研究发现：开放域聊天机器人在谈论事实相关话题以及观点相关话题时很容易产生前后不一致的现象。因此，本研究着重研究如何评估开放域聊天机器人前后事实和观点不一致的问题。针对此问题，受人类在对话中通过多次问询确认对话历史中事实信息的方式的启发，本研究认为是否可以正确回答关于对话历史的问题可以反映聊天机器人理解对话历史以及保持前后一致性的能力。基于此，本章提出了一种引入问询的对话前后一致性自动评测框架（AIH），根据对话前后一致性对不同的聊天机器人进行排名。

该框架基于聊天机器人-聊天机器人对话交互，并且在聊天机器人-聊天机器人对话交互过程中针对对话历史中出现的事实或观点信息进行再次问询，以此检测待测聊天机器人是否可以保持对话前后一致性。

接下来，本章将会详细介绍相关工作、AIH 框架设计，并深入分析 AIH 框架的有效性、高效性、稳定性，以及 AIH 框架中各模块的作用。

4.2　相关工作

本节将针对开放域对话系统评测中与本章相关的研究工作进行介绍。开放域对话系统前后一致性评测方法主要包含自动评测与人工评测。评测方法大多基于两类对话：第一类是静态对话，即基于已有的对话数据，检测对话回复是否与历史发生冲突；第二类是交互式对话，即在交互式对话过程中进行检测，适用于多数已有聊天机器人。

4.2.1　对话前后一致性静态评估

静态评估是指使用神经网络模型或人工评估基于静态前后生成的反应是否与预先定义的人物角色和对话历史相冲突。 (Welleck 等, 2019a; Song 等, 2020a) 关注与人物相关的一致性和与设定相关的一致性，并将聊天机器人前后一致性评价描述为自然语言推理问题。 Nie 等 (2020b) 构建了一个名为 DECODE 的新的人工标注数据集，并提出了一种基于结构化话语的方法来检测对话历史中的冲突。静态评估虽然具有成本效益，但并不能准确反映聊天机器人在现实世界中的会话能力。

4.2.2　对话前后一致性交互式评估

1. 人机交互式对话

为了追求更真实的评估，标准的方法是让人类与一个聊天机器人交谈，然后由上述模型或人类对其进行评估 (Mehri 和 Eskénazi, 2020c)。然而，除了收集人-机器对话的高成本之外，人类的认知压力也很高，这导致了不稳定的结果 (Dinan 等, 2020b)。

2. 聊天机器人—聊天机器人交互式对话

近年来，大大降低成本和降低人为偏见的聊天机器人—聊天机器人对话成为人们关注的焦点。 (Deriu 等, 2020) 提议人工对机器人的自言自语对话进行评

估，从而对聊天机器人的整体质量给出一个相对的排名。与这些方法不同的是，本研究关注聊天机器人的前后一致性，并插入问询来让聊天机器人重新回答观点或事实，同时引入了自动和人工两种方法来评估聊天机器人的前后一致性。

4.3 引入问询的开放域聊天机器人前后一致性检测框架

本节将对本研究所提出的对话前后一致性评估框架的技术细节进行介绍。如图4.2所示，本研究提出的 AIH 框架包含问询阶段（Inquiry Stage）和冲突检测阶段（Contradiction Recognition Stage），基于 Bot-Bot 对话交互。在问询阶段，本研究自动生成关于事实和观点的问题插入到当前 Bot-Bot 对话。在冲突检测阶段，本研究收集被测试聊天机器人针对问询阶段的问题的回复，使用自动评估模型或人工评估检测该回复是否与对应历史冲突。

在该框架中，共有 5 个模块：聊天机器人 1（Chatbot1），聊天机器人 2（Chatbot2），问询器（Inquirer），自动评估器（Auto Evaluator），以及人工评估器（Human Evaluator）。聊天机器人 1 和聊天机器人 2 是进行对话交互的两个聊天机器人，其中聊天机器人 2 是被测试聊天机器人。问询器负责提取对话中与事实和观点相关的实体，然后根据这些提取到的实体生成问题。自动评估器通常是一个冲突检测模型，用于自动检测聊天机器人 2 的回复是否与前后一致。人工评估器是人工评测，用于更加精确的评估。

具体地，有 N 个待测试聊天机器人 $\{B_1, B_2, ..., B_N\}$。对于每对聊天机器人(聊天机器人 1 和聊天机器人 2)，使聊天机器人 1 与聊天机器人 2 交互 K 轮。(1)在问询阶段，在聊天机器人 1 与聊天机器人 2 的对话过程中，对于聊天机器人 2 回复的每句话 u_{2k}，问询器提取该句话中事实和观点相关的实体，然后询问聊天机器人 2 一个与实体相关的问题 q_k，其中 k 是对话轮数。Chatbot2 回答问题 q_k 生成回复 r_k。当某句话中没有与事实或观点相关的实体时，不进行问题生成。(2)在冲突检测阶段，本研究使用神经网络模型（例如自然语言推理模型）或雇佣人工标注员进行判断对话语句对 $\{u_{2k}, r_k\}$ 是否存在冲突的现象。在 AIH 框架下，针对每个聊天机器人对，们收集至少 M 个对话，根据这些对话计算每个聊天机器人的前后一致性程度，据此对聊天机器人进行排名。通过这种方式，可以有效高效地评测聊天机器人的前后一致性。接下来将介绍问询阶段以及冲突检测阶段的具体细节。

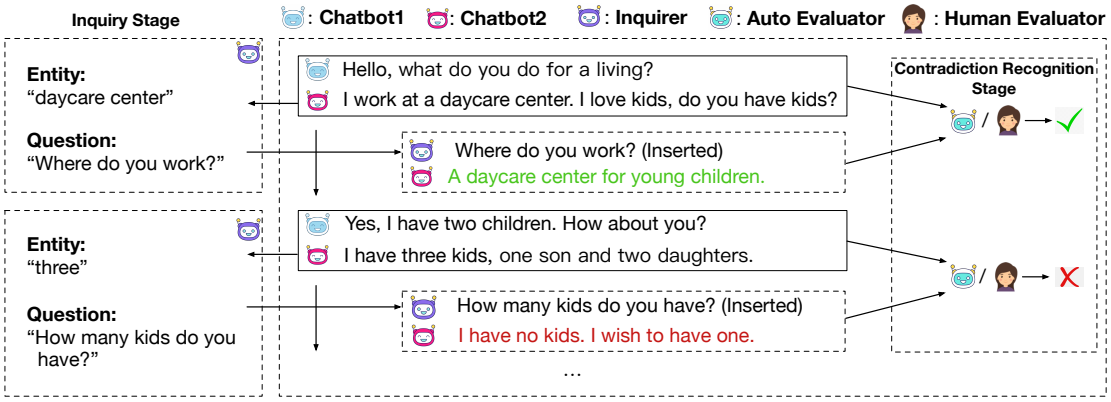

图 4.2 引入问询的开放域聊天机器人前后一致性检测框架

Figure 4.2 Overview of Addressing Inquiries about History framework for evaluating open-domain chatbot consistency.

4.3.1　问询阶段

根据本研究过程的观察和之前的研究工作 (Nie 等, 2020a)，在通常的人机对话或 Bot-Bot 对话中，当谈论重复的事实和观点时，尤其是在类似的问题之后，矛盾更容易发生。因此，为了模拟这种矛盾发生的过程，本研究让聊天机器人通过询问之前的事实和观点的相关问题来产生回复。在这种情况下，提出合适的问题是非常重要的。因此，首先从对话历史中提取有关事实和观点的实体，然后利用神经网络模型生成有关这些实体的问题。

1. 实体抽取

考虑到聊天机器人在谈论事实和观点时通常会产生矛盾，本研究使用 Stanza (Qi 等, 2020) 命名实体识别工具（一个流行的自然语言分析包），从包含人、组织、位置等[2]的对话语句中提取命名实体。例如，对于 "我明年想去纽约。"，该工具可以提取出两个实体："纽约" 和 "明年"。

2. 问题生成模型

对于问题生成，该框架使用 UniLM (Dong 等, 2019) 模型，该模型在 SQuAD 数据集 (Rajpurkar 等, 2016) 上进行问题生成任务 (Wang 等, 2020) 的微调。AIH 框架中使用了公开的模型[3]，给定在之前提取到的实体，UniLM 为每个实体生成一个合适的问题。例如，给出 "纽约" 和 "明年我想去纽约。"，模型会生成 "你明年想去哪里？" 然后 AIH 框架随机选择一个问题，把它插入 bot-bot 对话中。

[2]共有 18 种类型的实体，详细信息请见 (Weischedel 等, 2013)。

[3]https://github.com/artitw/text2text

4.3.2 冲突检测阶段

在 AIH 框架中，因为问题 q_k 基于聊天机器人 2 的对话语句 u_{2k}，所以聊天机器人 2 的回复应该与对话语句 u_{2k} 保持一致。因此，自动评估器和人工评估器只需要考虑回答 r_k 与对话语句 u_{2k} 是否一致。

1. 自动评估器

对于自动评价，自动评估器一般是冲突检测模型。自动评估器将聊天机器人 2 回答的 r_k 和之前的对话语句 u_{2k} 作为输入，输出冲突分数 y_k。可以表示为:

$$y_k = f_\theta(r_k, u_{2k}), \tag{4.1}$$

其中 f_θ 为冲突检测函数，θ 为参数。与其他考虑整个对话的冲突检测方法相比，该自动评估器可以避免整个对话所包含的很多不存在冲突的对话语句带来的噪声。本实验选择在多领域自然语言推理数据集 (Williams 等, 2018) 上进行微调的 Roberta-large 模型 (Liu 等, 2019) 作为前后冲突自动检测器的实现[4] 。

2. 人工评估器

在传统的对话一致性评价方法中，人工标注员被要求阅读整个对话，并给出一个整体上的一致性评分，通常为 0 或 1 分。这些方法的成本高，标注员间一致性低，并且没有具体的标注标准，而且人工标注员很难给整个对话打分 (Mehri 和 Eskénazi, 2020c)。在 AIH 框架中，人工评估者只被要求判断聊天机器人 2 的回答是否与之前的对话语句 u_{2k} 一致，这比传统方法更具体、更简单。从而降低了评价成本，提高了评价质量。另外，AIH 框架中的人工标注比传统的方法精细得多，可以为对话系统的开发周期提供更多的评测信息。

4.3.3 对话前后一致性指标及排序方式

根据前面的结果，本实验可以得到一个不同聊天机器人在前后一致性上的排名列表。形式上，对于每一对聊天机器人 $\{B_i, B_j\}$，收集 M 个对话。对于每个问询-回答对，通过将 y_k 与阈值 τ 进行比较来检测是否冲突:

$$c_k = \mathbb{I}(f_\theta(r_k, u_{2k}) > \tau). \tag{4.2}$$

聊天机器人 B_j 在聊天机器人对 $\{B_i, B_j\}$ 中的前后冲突率可以表示为:

$$C_{ij} = \frac{1}{M}\sum^{m} c_k, \tag{4.3}$$

[4] https://huggingface.co/roberta-large-mnli

其中，m 是每个对话中问询的数量，M 是总体的问询-回答对。聊天机器人 B_j 的整体前后冲突率可以被表示为：

$$C_j = \frac{1}{N}\sum_{i=1}^{N} C_{ij}. \tag{4.4}$$

最终，AIH 框架使用整体前后冲突率对聊天机器人进行排名。

4.4　实验结果与分析

为了验证所提出的 AIH 框架的有效性与高效性，本文开展了一系列的实验。本文将先详细介绍实验所使用的聊天机器人，然后介绍实验中的详细设置，最后本文针对框架的有效性、高效性、稳定性进行了充分的实验。

4.4.1　聊天机器人选择

本研究选择了四个常用的开放域聊天机器人进行实验。

- Blender (Smith 等, 2020b)：首先在 Reddit 数据集 (Baumgartner 等, 2020) 上进行预训练，然后使用高质量的人类标注对话数据集 (BST) 进行微调，该数据集包含四个数据集：混合技能谈话 (Smith 等, 2020b)、维基百科向导 (Dinan 等, 2019)、ConvAI2 (Dinan 等, 2020b) 和共情对话 (Rashkin 等, 2019)。通过微调，Blender 可以学习融合参与、知识、同理心和个性的会话技能。Blender 有三种模型尺寸:90M, 2.7B 和 9.4B。由于 2.7B 参数模型在 (Smith 等, 2020b) 中获得了最好的性能，在本实验中使用了 2.7B 版本。
- Plato-2 (Bao 等, 2020a) 是一个开放域聊天机器人，在 Reddit 数据集上进行预训练，并与 BST 数据集进行微调，据称比 Blender 更优秀。根据 (Bao 等, 2020a) 的描述，在本实验中选择了 1.6B 参数版本。
- DialoGPT (Zhang 等, 2020a) 是在 GPT-2 (Radford 等, 2019) 基础上使用 Reddit 评论进行预训练的。DialoGPT 有三种模型尺寸：117M、345M 和 762M。本实验在 BST 数据集上对 762M 版本进行了微调。
- DialoFlow (Li 等, 2021c,f) 是 DSTC9 交互式对话评估比赛 (Gunasekara 等, 2021) 中的一种方法。本实验基于 GPT2-large (Radford 等, 2019) 复现了 DialoFlow 模型，并使用 BST 数据集对其进行了微调。

4.4.2 实验设置

本实验采用四种实验范式来评估 AIH 框架的有效性和高效性。

（1）聊天机器人-聊天机器人（Bot-Bot）交互。对于 Bot-Bot 交互，最大交互回合设置为 15。所有聊天机器人都利用 Nucleus Sampling (Holtzman 等, 2020) 生成响应时的 $p = 0.9$。对于每对聊天机器人，收集至少 200 个对话。

（2）人工标注。为了验证框架的有效性，实验中进行了人工评估。对于 AIH 框架下的 Bot-Bot 对话，本研究聘请了来自一家商业数据注释公司的三位专业人工标注员分别标注三个问题：询问机器人是否生成合适的问题，聊天机器人是否对问题进行了相关的回答，以及聊天机器人的回答是否与对话历史发生冲突。对于每对聊天机器人，本实验随机抽取 50 个对话进行标注，通过投票计算出最终的评分。

在人机自然交互和专家评估中，本实验将四个聊天机器人部署在远程服务器上，并设计了一个网页界面。人类每次都可以通过网页界面与随机的聊天机器人聊天，并给出前后一致性评分，但不知道他们正在与哪一个聊天机器人聊天。

（3）人机对话交互。对于每个聊天机器人，过滤掉小于 5 回合的对话和带有辱骂性语言的对话。对于每个聊天机器人，至少有 40 个符合条件的对话。然后本研究雇佣三个专业的人工标注员来分别标注来自聊天机器人的每个对话语句是否一致。

（4）专家评估。为了获得聊天机器人前后一致性的专家排名，本研究从实验室邀请了 3 名有 2-3 年对话系统开发经验的专家志愿者，与每个机器人聊天至少 10 次，每次约 15 轮。在聊天过程中，专家被要求有意地诱导聊天机器人重新回答有关对话历史的问题，并将一致性评分从 0 分到 1 分。

正式评估前，专家需要与每个聊天机器人至少交谈 20 次。三位专家的得分的平均值作为整体前后一致性得分。在自动评估之前，专家评估和人工标注是完成的。自动评估后进行人机交互。所有的人工评估都独立于自动评估。

4.4.3 有效性实验结果

自动评测框架的有效性是设计评测方式的第一原则，本小节展示了在 AIH 框架下的专家评估、自动评估和人工评估中的排名结果，证明 AIH 框架的有效性。

1. 专家评估排名

表4.1显示了不同聊天机器人的专家一致性得分。我们可以发现 Blender 的专家一致性得分最高，达到了 0.85。这四个聊天机器人的一致性排名是：Plato > DialoGPT > DialoFlow > Blender，可以作为标准参考。

2. 自动评估结果

表4.2显示了每对聊天机器人在自动评估中的冲突率。冲突率越低，一致性越好。列名和行名分别表示聊天机器人 1 和聊天机器人 2。列名中的 “Avg.” 表示每个聊天机器人的整体冲突率。行名中的 “Avg.” 可以被视为诱导其他聊天机器人产生相互冲突的事实或观点的能力。在自动评估中，聊天机器人的一致性排名为 Plato > DialoGPT > DialoFlow > Blender，与专家评估相同。Blender 发生冲突率最高。

3. 人工评估结果

我们在表4.2的底部列出了人工评估结果。Blender 获得了最高的冲突率。同时，人工评估结果也提供了同样的前后一致性排名：Plato > DialoGPT > DialoFlow > Blender，和专家评估排名一致。

在我们的框架中，自动评估和人工评估都能给出与专家评估相同的前后一致性排名，这表明我们的 AIH 框架具有通用性，能够有效地评估聊天机器人的前后一致性。

表 4.1 对话聊天机器人的专家评估一致性分数

Table 4.1 Expert consistency score for each chatbot

Expert Consistency Score ↑				
	Blender	**Plato-2**	**DialoGPT**	**DialoFlow**
Expert.1	0.55	0.80	0.72	0.69
Expert.2	0.37	0.87	0.60	0.56
Expert.3	0.31	0.89	0.60	0.55
Avg.	0.41	**0.85**	0.64	0.60

表 4.2 任意两个聊天机器人对的人工评估冲突率与自动评估冲突率

Table 4.2 Auto evaluation and human evaluation contradiction rate for each chatbot pair

Contradiction Rate (Auto $\tau = 0.15$) ↓					
	Blender	**Plato-2**	**DialoGPT**	**DialoFlow**	**Avg.**
Blender	0.431	0.240	0.324	0.362	0.339
Plato-2	0.431	0.263	0.293	0.357	0.336
DialoGPT	0.425	0.251	0.344	0.345	0.341
DialoFlow	0.427	0.264	0.344	0.371	0.351
Avg.	0.428	**0.255**	0.326	0.359	0.342
Contradiction Rate (Human) ↓					
	Blender	**Plato-2**	**DialoGPT**	**DialoFlow**	**Avg.**
Blender	0.487	0.282	0.398	0.396	0.391
Plato-2	0.411	0.212	0.500	0.435	0.390
DialoGPT	0.404	0.211	0.304	0.431	0.338
DialoFlow	0.462	0.268	0.310	0.377	0.354
Avg.	0.441	**0.243**	0.378	0.410	0.368

4.4.4 高效性实验结果

以往的人机对话交互前后一致性评价方法成本高、耗时长，严重延缓了对话系统的开发周期。在这一节中，我们试图说明我们所提出的关于历史的问询 AIH 框架与其他方法相比，具有时间和成本效益，并且可以加速对话系统的提升进程。如表4.3所示，我们比较了两个方面的时间成本：（1）收集对话的时间；（2）发现对话中的冲突的时间。AIH 框架基于 Bot-Bot 对话，因此可以忽略收集对话的时间，而人机对话每次对话大约需要 4 分钟。在冲突检测时间方面，以往的方法考虑到整个对话需要 1 分钟左右的时间，而在我们提出的框架中，人工标注只需 24 秒左右，或者使用自动评价，时间忽略不计。除此之外，我们还比较了每次对话中出现冲突的次数。如表4.3所示，在 AIH 框架中，聊天机器人产生的冲突比之前的方法多得多。检测到的冲突有助于聊天机器人开发者进一步提高聊天机器人的一致性。

表 4.3 AIH 框架高效性实验结果

Table 4.3 Effectiveness of the AIH framework

Method	Time (Sec)	Contradiction
AIH (Auto)	- + -	1.56
AIH (Human)	- + 24	1.69
Human-bot	246 + 59	0.50

与以前的方法相比，我们提出的 AIH 框架可以在更短的时间内检测出更多的冲突，同时也有效的评估了聊天机器人的前后一致性。因此，AIH 框架将加速聊天机器人前后一致性的提升进程。

4.4.5　稳定性实验结果

评估框架的一个关键需求是，评估过程的重复执行会产生相同的结果。本实验测量每对聊天机器人之间需要进行多少次对话才能保证一个稳定的排名。

本实验对每对聊天机器人随机抽样 $\hat{S}$ 会话，使用自动评估计算前后一致性排名，其中 $\hat{S} \in \{1, \cdots, 200\}$。本实验将此子抽样过程重复 1000 次，并计算得到与专家评估排序相同的排序的准确性。如图4.3所示，当 $\hat{S} > 100$ 时，四个聊天机器人在 95% 的情况下与专家的排名结果相同，保证了一个稳定的排名。本研究也做了更深入的分析。排名的稳定性对该排名方式的有效性意义重大。表4.1显示 DialoGPT 和 DialoFlow 的一致性得分比较接近。本实验进行了稳定性分析，去掉了一个聊天机器人。图4.3显示当在 DialoGPT 或 DialoFlow 之间留下一个时，$\hat{S} = 50$ 个对话框实现了稳定性。

在 AIH 框架中，稳定评估所需的对话数量取决于要测试的聊天机器人，而更多的对话通常会导致更稳定的评估。通常情况下，75 次对话就足以构成一次稳定有效的评估。

4.4.6　问题生成结果分析

由于在 AIH 框架下，问询阶段需要一个合适的问题，因此本小节深入分析了问询阶段问题生成的特点：问题数量与发生冲突情况、问题生成合理性。

1. 问题数量与发生冲突情况

本小节为每对聊天机器人随机抽样 200 个对话，并统计每对对话的问询回

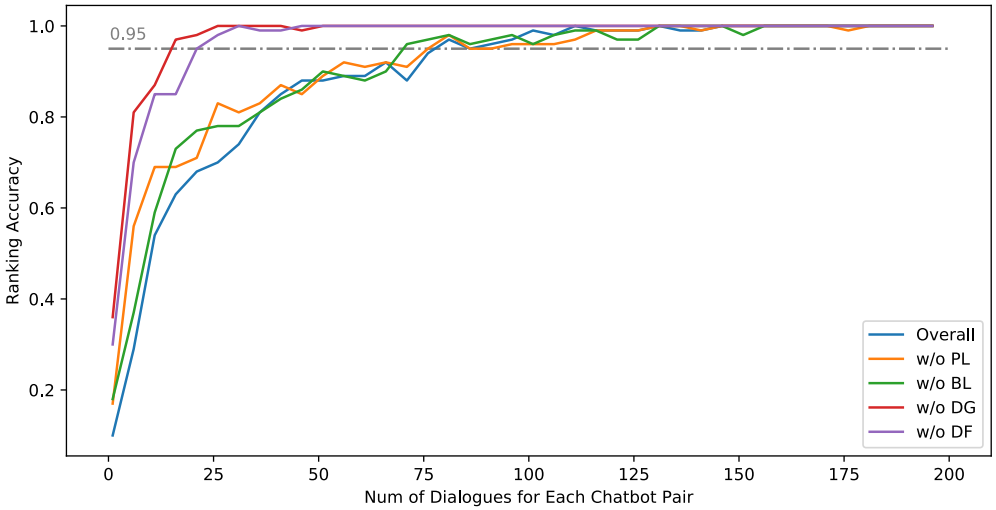

图 4.3 AIH 框架稳定性实验结果

Figure 4.3 Stability of AIH framework

答对的数量和发生冲突的平均数量。如表4.4所示，平均每个会话有 4.57 个问询回答对。当聊天机器人 Blender 分别作为待测聊天机器人和问询机器人时，每段对话有 6.37 对和 5.10 对问询回答对，这在所有聊天机器人中是最高的。问询回答对的数量表明，聊天机器人 Blender 可以更多地谈论角色和事实相关的话题，而聊天机器人 DialoGPT 很少提到这些事情。表4.4也显示了每个对话发生冲突的数量。类似地，聊天机器人 Blender 的对话中发生的冲突最多，也最有可能促使其他聊天机器人与它互动，发生与事实或观点相关的冲突。

2. 问题生成合理性

为了分析生成问题的合理性，本实验从每对聊天机器人中随机抽取 50 个对话，并让人工标注者根据所提供的前后来决定生成的问题是否合理，合理为 1 分，不合理为 0 分。如表4.5所示，总体合理性得分约为 0.93，说明问题生成策略简单有效。本小节进一步研究了错误的问题，发现大多数问题都是由于一般的问题生成模型不能很好地应用于对话场景。在开放领域的对话中设计更好的问题生成模型是未来的工作之一。

表 4.4 每两个聊天机器人对话中问询回答对数量及发生冲突的数量

Table 4.4 Amount of questions and contradictions in each chatbot pair

Number of Questions					
	Blender	**Plato-2**	**DialoGPT**	**DialoFlow**	**Avg.**
Blender	6.54	6.13	2.62	5.12	**5.10**
Plato-2	6.54	5.34	1.98	4.36	4.55
DialoGPT	6.25	4.45	1.67	3.79	4.04
DialoFlow	6.15	5.79	2.25	4.21	4.60
Avg.	**6.37**	5.42	2.13	4.37	4.57
Number of Contradictions ($\tau = 0.15$)					
	Blender	**Plato-2**	**DialoGPT**	**DialoFlow**	**Avg.**
Blender	2.61	1.28	1.61	1.50	**1.74**
Plato-2	2.82	1.40	0.58	1.56	1.53
DialoGPT	2.66	1.12	0.57	1.31	1.38
DialoFlow	2.63	1.53	0.77	1.56	1.61
Avg.	**2.73**	1.38	0.69	1.57	1.56

表 4.5 问题生成的合理性

Table 4.5 Appropriateness of question generation

Question Appropriateness					
	Blender	**Plato-2**	**DialoGPT**	**DialoFlow**	**Avg.**
Blender	0.932	0.960	0.922	0.936	0.938
Plato-2	0.942	0.976	0.940	0.948	0.951
DialoGPT	0.784	0.870	0.928	0.882	0.866
DialoFlow	0.867	0.934	0.922	0.939	0.915
Avg.	0.881	0.935	0.947	0.942	0.927

4.5 本章小结

本章针对开放域对话前后一致性自动评估难问题，提出了一种引入问询的开放域聊天机器人前后一致性评估框架——AIH。该框架的工作原理是，在聊天机器人与聊天机器人之间的对话中插入有关历史事实和观点的问题，并使用人工评估或神经网络模型来评估对话回复是否与对话历史一致。在此基础上，该框架可以对不同的聊天机器人在前后一致性上进行准确、高效的排序。实验结果表明，该框架能够有效地评估聊天机器人的上下一致性，评估结果与专家评估保持一致。此外，本研究的框架成本低和时间效率高，不仅可以给出整体一致性评分，而且可以提供准确的冲突位置，这可以加速聊天机器人的发展过程。

本工作只关注事实和观点相关的实体的冲突，未来的工作可以改进问询模块，探索更多种类的冲突。此外，未来的工作还应该开发一个更有效的开放域对话领域的冲突检测模块，而本工作只是利用通用的自然语言推理模型来检测冲突。目前开放域聊天机器人存在着严重的不一致性问题。本研究希望该工作能够促进并为未来开发前后一致性开放域聊天机器人的工作提供指导方针。

除此之外，开放域对话系统的前后文一致性提高有以下研究方向：1）利用本文提出的自动评估框架，搭建基于强化学习的反馈-优化系统，不断优化开放域对话系统的前后文一致性；2）收集大量的高质量对话数据，让对话系统在训练过程中，学会对话前后一致的关系；3）设计解码方法或后处理方法，从后处理角度杜绝前后文不一致现象的发生。

第 5 章　总结与展望

5.1　总结

本文受人类对话认知过程的启发，针对开放域对话系统中多轮对话逻辑不清、缺乏背景知识、前后一致性自动评估难等问题展开了深入的研究，并进行了以下三个方面的研究工作：

- 针对开放域对话中的多轮对话逻辑不清问题，本文受人类对话认知过程——“人类总是在继续对话之前考虑下一个对话回复对整体对话的影响” 的启发，在对对话历史建模过程中引入了高层语义信息流，充分考虑对话语句内的理解与对话语句间的语义逻辑关系。进一步设计了单向的流模块捕捉对话过程中的语义变化过程，提出了 DialoFlow 大规模对话预训练模型和基于 DialoFlow 模型的流评分交互式对话自动评测方法。对话回复生成与对话自动评估实验的实验结果证明本文基于高层语义信息流的对话历史建模方法的有效性，可以大大提高模型对对话历史的理解能力和开放域对话系统回复生成的逻辑性。

- 针对开放域对话中的缺乏背景知识问题，本文受人类会从周围获取多种模态的信息，辅助进行对话，并且多模态信息相互辅助理解的认知过程启发，提出了视听场景感知的统一的多模态 Transformer 模型。该模型以预训练的语言模型为基础，融入视觉、听觉、文本等多模态信息，通过多任务学习（对话回复生成建模，视频-音频序列建模，视频描述生成建模），促进多模态信息的理解，辅助开放域对话回复生成。本文通过实验验证了该模型的有效性，该模型可以生成流畅的、有信息量的、符合视听场景的对话回复。同时，本文也分析了该模型的不足之处，如多模态信息的基础表征能力限制该模型的多模态理解能力。

- 针对开放域对话前后一致性自动评估难的问题，本文受人类在对话中通过问询发现前后不一致的认知过程启发，提出了基于问询历史的开放域对话前后一致性评测框架——AIH。该框架通过聊天机器人-聊天机器人对话取代人机对话，并在对话过程中针对事实和观点相关的信息对被测聊天机器人进行问询，验证被测聊天机器人的回答是否与对话历史发生冲突。实验验证了该框架的有效性、高效性、稳定性。该框架大大提升了开放域对话前后一致性自动评估的效率，实现了稳定的、与专家评估一致的评估效果，也为之后开放域对话一致性的

提高提供了自动评测的工具。

5.2 展望

科学的研究是一种循序渐进的过程，离不开大量科学研究人员不断的探索和归纳。开放域对话系统经历了专家系统、基于统计的开放域对话系统和基于深度神经网络的开放域对话系统的发展历程。近年来，开放域对话系统取得了令人欣喜的进展，但作为该领域的研究人员我们仍需清晰认识到，现在的开放域对话系统还有很多不足之处，存在一些亟待解决的问题，本文仅对开放域对话系统缺乏知识、对话逻辑不清、对话前后一致性自动评估难问题进行了探究，并给出了对应的解决方案。但是这些解决方案并不能完全解决这一类的问题，这些问题需要接下来更多的研究推进解决。从人类认知学上讲，开放域对话系统本质上是对人类对话认知过程的模拟，更深层次理解人类是如何进行对话的，对开放域对话系统的发展至关重要。从开放域对话系统技术上讲，开放域对话系统应该模拟人类对话认知过程，如从外界实时获取需要的知识，从自身记忆获取需要的知识，更深层次理解对话的逻辑，有一个实时有效的评估反馈系统，将是未来开放域对话研究的重要问题。

参考文献

ADIWARDANA D, LUONG M T, SO D R, et al. Towards a human-like open-domain chatbot[J]. ArXiv, 2020, abs/2001.09977.

ADIWARDANA D, LUONG M T, SO D R, et al. Towards a human-like open-domain chatbot[J]. arXiv preprint arXiv:2001.09977, 2020.

AGRAWAL A, LU J, ANTOL S, et al. Vqa: Visual question answering[J]. International Journal of Computer Vision, 2017, 123(1):4-31.

ALAMRI H, CARTILLIER V, LOPES R G, et al. Audio visual scene-aware dialog (AVSD) challenge at DSTC7[J]. CoRR, 2018, abs/1806.00525.

BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

BAO S, HE H, WANG F, et al. Plato-2: Towards building an open-domain chatbot via curriculum learning[J]. ArXiv, 2020, abs/2006.16779.

BAO S, HE H, WANG F, et al. PLATO: Pre-trained dialogue generation model with discrete latent variable[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020b: 85-96.

BAO S, HE H, WANG F, et al. PLATO-2: towards building an open-domain chatbot via curriculum learning[J/OL]. CoRR, 2020, abs/2006.16779. https://arxiv.org/abs/2006.16779.

BAUMGARTNER J, ZANNETTOU S, KEEGAN B, et al. The pushshift reddit dataset[C/OL]// CHOUDHURY M D, CHUNARA R, CULOTTA A, et al. Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020. AAAI Press, 2020: 830-839. https://aaai.org/ojs/index.php/ICWSM/article/view/7347.

BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C/OL]// LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. 2020a. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C/OL]// LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. 2020b. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

BROWN-SCHMIDT S, KONOPKA A E. Processes of incremental message planning during conversation[J/OL]. Psychonomic bulletin & review, 2015, 22(3):833-843. https://link.springer.com/article/10.3758%2Fs13423-014-0714-2.

CHEN Y, LI L, YU L, et al. UNITER: universal image-text representation learning[C]//Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020: volume 12375. 2020: 104-120.

CHU Y, LIN K, HSU C, et al. Multi-step joint-modality attention network for scene-aware dialogue system[J]. CoRR, 2020, abs/2001.06206.

CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[Z]. 2014.

DAS A, KOTTUR S, GUPTA K, et al. Visual dialog[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 326-335.

DENKOWSKI M J, LAVIE A. Meteor universal: Language specific translation evaluation for any target language[C]//Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA. 2014: 376-380.

DERIU J, TUGGENER D, VON DÄNIKEN P, et al. Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems[C/OL]//WEBBER B, COHN T, HE Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, 2020: 3971-3984. https://doi.org/10.18653/v1/2020.emnlp-main.326.

DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019a: 4171-4186.

DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C/OL]//BURSTEIN J, DORAN C, SOLORIO T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 2019b: 4171-4186. https://doi.org/10.18653/v1/n19-1423.

DINAN E, ROLLER S, SHUSTER K, et al. Wizard of wikipedia: Knowledge-powered conversational agents[C/OL]//7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019. https://openreview.net/forum?id=r1l73iRqKm.

DINAN E, FAN A, WILLIAMS A, et al. Queens are powerful too: Mitigating gender bias in

dialogue generation[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020a: 8173-8188.

DINAN E, LOGACHEVA V, MALYKH V, et al. The second conversational intelligence challenge (convai2)[M/OL]//The NeurIPS'18 Competition. Springer, 2020b: 187-208. http://arxiv.org/abs/1902.00098.

DONG L, YANG N, WANG W, et al. Unified language model pre-training for natural language understanding and generation[C/OL]//WALLACH H M, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. 2019: 13042-13054. https://proceedings.neurips.cc/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html.

FEI Z, LI Z, ZHANG J, et al. Towards expressive communication with internet memes: A new multimodal conversation dataset and benchmark[J/OL]. CoRR, 2021, abs/2109.01839. https://arxiv.org/abs/2109.01839.

GAO X, ZHANG Y, GALLEY M, et al. Dialogue response ranking training with large-scale human feedback data[C/OL]//WEBBER B, COHN T, HE Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, 2020: 386-395. https://doi.org/10.18653/v1/2020.emnlp-main.28.

GENG S, GAO P, HORI C, et al. Spatio-temporal scene graphs for video dialog[J]. CoRR, 2020, abs/2007.03848.

GHAZVININEJAD M, BROCKETT C, CHANG M W, et al. A knowledge-grounded neural conversation model[C]//AAAI. 2018.

GOYAL Y, KHOT T, SUMMERS-STAY D, et al. Making the v in vqa matter: Elevating the role of image understanding in visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6904-6913.

GU J C, LI T, LIU Q, et al. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots[J]. Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020.

GU X, YOO K M, HA J. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances[J/OL]. CoRR, 2020, abs/2012.01775. https://arxiv.org/abs/2012.01775.

GU X, YOO K M, HA J. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances[C/OL]//Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Vir-

tual Event, February 2-9, 2021. AAAI Press, 2021: 12911-12919. https://ojs.aaai.org/index.php/AAAI/article/view/17527.

GUNASEKARA C, KIM S, D'HARO L F, et al. Overview of the ninth dialog system technology challenge: Dstc9[J/OL]. Proceedings of the 9th Dialog System Technology Challenge Workshop in AAAI2021, 2021. https://arxiv.org/abs/2011.06486.

GUPTA P, MEHRI S, ZHAO T, et al. Investigating evaluation of open-domain dialogue systems with human generated multiple references[C/OL]//NAKAMURA S, GASIC M, ZUCKERMAN I, et al. Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019. Association for Computational Linguistics, 2019: 379-391. https://doi.org/10.18653/v1/W19-5944.

HAN S, BANG J, RYU S, et al. Exploiting knowledge base to generate responses for natural language dialog listening agents[C]//Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Prague, Czech Republic: Association for Computational Linguistics, 2015: 129-133.

HERSHEY S, CHAUDHURI S, ELLIS D P, et al. Cnn architectures for large-scale audio classification[C]//2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017: 131-135.

HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. 1997:1735-1780.

HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration[C/OL]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. https://openreview.net/forum?id=rygGQyrFvH.

HORI C, ALAMRI H, WANG J, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 2352-2356.

HU B, LU Z, LI H, et al. Convolutional neural network architectures for matching natural language sentences[C]//Advances in Neural Information Processing Systems. 2014.

INABA M, TAKAHASHI K. Neural utterance ranking model for conversational dialogue systems [C]//Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Los Angeles: Association for Computational Linguistics, 2016: 393-403.

KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[J]. CoRR, 2017, abs/1705.06950.

KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[J]. arXiv preprint arXiv:1705.06950, 2017.

KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional

neural networks[C]//PEREIRA F, BURGES C J C, BOTTOU L, et al. Advances in Neural Information Processing Systems. Curran Associates, Inc., 2012.

LAVIE A, AGARWAL A. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments[C/OL]//CALLISON-BURCH C, KOEHN P, FORDYCE C S, et al. Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007. Association for Computational Linguistics, 2007: 228-231. https://www.aclweb.org/anthology/W07-0734/.

LE H, SAHOO D, CHEN N, et al. Multimodal transformer networks for end-to-end video-grounded dialogue systems[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 5612-5623.

LEE H, YOON S, DERNONCOURT F, et al. DSTC8-AVSD: multimodal semantic transformer network with retrieval style word generator[J]. CoRR, 2020, abs/2004.08299.

LI J, GALLEY M, BROCKETT C, et al. A persona-based neural conversation model[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 994-1003.

LI J, LIU C, TAO C, et al. Dialogue history matters! personalized response selectionin multi-turn retrieval-based chatbots[J]. CoRR, 2021.

LI Y, SU H, SHEN X, et al. Dailydialog: A manually labelled multi-turn dialogue dataset[C/OL]// KONDRAK G, WATANABE T. Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers. Asian Federation of Natural Language Processing, 2017: 986-995. https://www.aclweb.org/anthology/I17-1099/.

LI Z, NIU C, MENG F, et al. Incremental transformer with deliberation decoder for document grounded conversations[C/OL]//KORHONEN A, TRAUM D R, MÀRQUEZ L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019a: 12-21. https://doi.org/10.18653/v1/p19-1002.

LI Z, NIU C, MENG F, et al. Incremental transformer with deliberation decoder for document grounded conversations[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. 2019b: 12-21.

LI Z, LI Z, ZHANG J, et al. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog[J/OL]. IEEE ACM Trans. Audio Speech Lang. Process., 2021, 29: 2476-2483. https://doi.org/10.1109/TASLP.2021.3065823.

LI Z, LI Z, ZHANG J, et al. Wechat ai's submission for dstc9 interactive dialogue evaluation track [J/OL]. Proceedings of the 9th Dialog System Technology Challenge Workshop in AAAI2021, 2021. https://arxiv.org/abs/2101.07947.

LI Z, ZHANG J, FEI Z, et al. Addressing inquiries about history: An efficient and practical framework for evaluating open-domain chatbot consistency[C/OL]//ZONG C, XIA F, LI W, et al. Findings of ACL: ACL/IJCNLP 2021 Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021. Association for Computational Linguistics, 2021d: 1057-1067. https://doi.org/10.18653/v1/2021.findings-acl.91.

LI Z, ZHANG J, FEI Z, et al. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances[J]. CoRR, 2021.

LI Z, ZHANG J, FEI Z, et al. Conversations are not flat: Modeling the intrinsic information flow between dialogue utterances[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. 2021f.

LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.

LIN C Y, OCH F J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics[C/OL]//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). 2004: 605-612. https://www.aclweb.org/anthology/P04-1077/.

LIU Q, CHEN Y, CHEN B, et al. You impress me: Dialogue generation via mutual persona perception[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020a: 1417-1427.

LIU S, CHEN H, REN Z, et al. Knowledge diffusion for neural dialogue generation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018: 1489-1498.

LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized BERT pretraining approach[J/OL]. CoRR, 2019, abs/1907.11692. http://arxiv.org/abs/1907.11692.

LIU Y, OTT M, GOYAL N, et al. Ro{bert}a: A robustly optimized {bert} pretraining approach[Z]. 2020b.

LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C/OL]//7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019. https://openreview.net/forum?id=Bkg6RiCqY7.

LOWE R, POW N, SERBAN I, et al. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems[C]//Proceedings of the 16th Annual Meeting of the

Special Interest Group on Discourse and Dialogue. Prague, Czech Republic: Association for Computational Linguistics, 2015: 285-294.

LU J, ZHANG C, XIE Z, et al. Constructing interpretive spatio-temporal features for multi-turn responses selection[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 44-50.

LU J, REN X, REN Y, et al. Improving contextual language models for response retrieval in multi-turn conversation[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: Association for Computing Machinery, 2020: 1805–1808.

MADOTTO A, WU C, FUNG P. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. 2018: 1468-1478.

MEHRI S, ESKÉNAZI M. USR: an unsupervised and reference free evaluation metric for dialog generation[C/OL]//JURAFSKY D, CHAI J, SCHLUTER N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, 2020a: 681-707. https://doi.org/10.18653/v1/2020.acl-main.64.

MEHRI S, ESKÉNAZI M. Unsupervised evaluation of interactive dialog with dialogpt[C/OL]// PIETQUIN O, MURESAN S, CHEN V, et al. Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020. Association for Computational Linguistics, 2020b: 225-235. https://www.aclweb.org/anthology/2020.sigdial-1.28/.

MEHRI S, ESKÉNAZI M. Unsupervised evaluation of interactive dialog with dialogpt[C/OL]// PIETQUIN O, MURESAN S, CHEN V, et al. Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020. Association for Computational Linguistics, 2020c: 225-235. https://www.aclweb.org/anthology/2020.sigdial-1.28/.

MIKOLOV T, KARAFIáT M, BURGET L, et al. Recurrent neural network based language model.[C]//INTERSPEECH. 2010.

NGUYEN D T, SHARMA S, SCHULZ H, et al. From film to video: Multi-turn question answering with multi-modal context[J]. CoRR, 2018, abs/1812.07023.

NIE Y, WILLIAMSON M, BANSAL M, et al. I like fish, especially dolphins: Addressing contradictions in dialogue modelling[Z]. 2020a.

NIE Y, WILLIAMSON M, BANSAL M, et al. I like fish, especially dolphins: Addressing contra-

dictions in dialogue modeling[J/OL]. CoRR, 2020, abs/2012.13391. https://arxiv.org/abs/2012.13391.

PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C/OL]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. ACL, 2002a: 311-318. https://www.aclweb.org/anthology/P02-1040/.

PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002b: 311-318.

PASUNURU R, BANSAL M. Dstc7-avsd: Scene-aware video-dialogue systems with dual attention [C]//DSTC7 at AAAI2019 Workshop. 2019.

QI P, ZHANG Y, ZHANG Y, et al. Stanza: A Python natural language processing toolkit for many human languages[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020. https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

QIAN Q, HUANG M, ZHAO H, et al. Assigning personality/profile to a chatting machine for coherent conversation generation[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. International Joint Conferences on Artificial Intelligence Organization, 2018: 4279-4285.

RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8):9.

RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020:1-67.

RAJPURKAR P, ZHANG J, LOPYREV K, et al. Squad: 100, 000+ questions for machine comprehension of text[C/OL]//SU J, CARRERAS X, DUH K. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. The Association for Computational Linguistics, 2016: 2383-2392. https://doi.org/10.18653/v1/d16-1264.

RASHKIN H, SMITH E M, LI M, et al. Towards empathetic open-domain conversation models: A new benchmark and dataset[C/OL]//KORHONEN A, TRAUM D R, MÀRQUEZ L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019: 5370-5381. https://doi.org/10.18653/v1/p19-1534.

REDDY S, CHEN D, MANNING C D. Coqa: A conversational question answering challenge[J]. Transactions of the Association for Computational Linguistics, 2019, 7:249-266.

SANABRIA R, PALASKAR S, METZE F. Cmu sinbad’s submission for the dstc7 avsd challenge [C]//DSTC7 at AAAI2019 Workshop. 2019.

SANKAR C, SUBRAMANIAN S, PAL C, et al. Do neural dialog systems use the conversation history effectively? an empirical study[C/OL]//KORHONEN A, TRAUM D R, MÀRQUEZ L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019: 32-37. https://doi.org/10.18653/v1/p19-1004.

SERBAN I, KLINGER T, TESAURO G, et al. Multiresolution recurrent neural networks: An application to dialogue response generation[C]//AAAI. 2017.

SERBAN I V, SORDONI A, BENGIO Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models[C/OL]//SCHUURMANS D, WELLMAN M P. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA. AAAI Press, 2016a: 3776-3784. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957.

SERBAN I V, SORDONI A, BENGIO Y, et al. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models[C]//AAAI Press, 2016b: 3776-3783.

SERBAN I V, SORDONI A, LOWE R, et al. A hierarchical latent variable encoder-decoder model for generating dialogues.[Z]. 2016c.

SHAN Y, LI Z, ZHANG J, et al. A contextual hierarchical attention network with adaptive objective for dialogue state tracking[C/OL]//JURAFSKY D, CHAI J, SCHLUTER N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, 2020: 6322-6333. https://doi.org/10.18653/v1/2020.acl-main.563.

SHANG L, LU Z, LI H. Neural responding machine for short-text conversation[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1577-1586.

SHOEYBI M, PATWARY M, PURI R, et al. Megatron-lm: Training multi-billion parameter language models using model parallelism[J/OL]. arXiv preprint arXiv:1909.08053, 2019. http://arxiv.org/abs/1909.08053.

SMITH E M, WILLIAMSON M, SHUSTER K, et al. Can you put it all together: Evaluating conversational agents’ ability to blend skills[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020a: 2021-2030.

SMITH E M, WILLIAMSON M, SHUSTER K, et al. Can you put it all together: Evaluating conversational agents’ ability to blend skills[C/OL]//JURAFSKY D, CHAI J, SCHLUTER N,

et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, 2020b: 2021-2030. https://doi.org/10.18653/v1/2020.acl-main.183.

SONG H, WANG Y, ZHANG W N, et al. Profile consistency identification for open-domain dialogue agents[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020a: 6651-6662.

SONG H, WANG Y, ZHANG W, et al. Profile consistency identification for open-domain dialogue agents[C/OL]//WEBBER B, COHN T, HE Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, 2020b: 6651-6662. https://doi.org/10.18653/v1/2020.emnlp-main.539.

SORDONI A, GALLEY M, AULI M, et al. A neural network approach to context-sensitive generation of conversational responses[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 196-205.

SUN C, MYERS A, VONDRICK C, et al. Videobert: A joint model for video and language representation learning[C]//2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. 2019: 7463-7472.

TAO C, WU W, XU C, et al. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 1-11.

TIAN Z, YAN R, MOU L, et al. How to make context more useful? an empirical study on context-aware neural conversational models[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 231-236.

URBANEK J, FAN A, KARAMCHETI S, et al. Learning to speak and act in a fantasy text adventure game[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. 2019: 673-683.

VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//GUYON I, LUXBURG U V, BENGIO S, et al. Advances in Neural Information Processing Systems. 2017a.

VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017b: 5998-6008.

VEDANTAM R, LAWRENCE ZITNICK C, PARIKH D. Cider: Consensus-based image descrip-

tion evaluation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4566-4575.

WANG J, WANG X, LI F, et al. Group linguistic bias aware neural response generation[C]// Proceedings of the 9th SIGHAN Workshop on Chinese Language Processing. Association for Computational Linguistics, 2017: 1-10.

WANG W, HOI S C, JOTY S. Response selection for multi-party conversations with dynamic topic tracking[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6581-6591.

WEISCHEDEL R, PALMER M, MARCUS M, et al. Ontonotes release 5.0 ldc2013t19[J]. Linguistic Data Consortium, Philadelphia, PA, 2013, 23.

WELLECK S, WESTON J, SZLAM A, et al. Dialogue natural language inference[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019a: 3731-3741.

WELLECK S, WESTON J, SZLAM A, et al. Dialogue natural language inference[C/OL]// KORHONEN A, TRAUM D R, MÀRQUEZ L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019b: 3731-3741. https://doi.org/10.18653/v1/p19-1363.

WILLIAMS A, NANGIA N, BOWMAN S. A broad-coverage challenge corpus for sentence understanding through inference[C/OL]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, 2018: 1112-1122. http://aclweb.org/anthology/N18-1101.

WOLF T, DEBUT L, SANH V, et al. Huggingface's transformers: State-of-the-art natural language processing[J]. CoRR, 2019, abs/1910.03771.

WOLF T, SANH V, CHAUMOND J, et al. Transfertransfo: A transfer learning approach for neural network based conversational agents[J]. CoRR, 2019, abs/1901.08149.

WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv:1609.08144, 2016.

WU Y, WU W, XING C, et al. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 496-505.

XING C, WU W, WU Y, et al. Hierarchical recurrent attention network for response generation.[J]. CoRR, 2017.

YAN R, SONG Y, WU H. Learning to respond with deep neural networks for retrieval-based human-

computer conversation system[C]//New York, NY, USA: Association for Computing Machinery, 2016: 55–64.

YANG M, ZHAO Z, ZHAO W, et al. Personalized response generation via domain adaptation[J]. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017.

YOUNG T, CAMBRIA E, CHATURVEDI I, et al. Augmenting end-to-end dialogue systems with commonsense knowledge[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32.

YUAN C, ZHOU W, LI M, et al. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 111-120.

ZHANG H, CAI J, XU J, et al. Pretraining-based natural language generation for text summarization [C]//Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019. 2019a: 789-797.

ZHANG S, DINAN E, URBANEK J, et al. Personalizing dialogue agents: I have a dog, do you have pets too?[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018a: 2204-2213.

ZHANG W, HU J, FENG Y, et al. Refining source representations with relation networks for neural machine translation[C/OL]//BENDER E M, DERCZYNSKI L, ISABELLE P. Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. Association for Computational Linguistics, 2018b: 1292-1303. https://aclanthology.org/C18-1110/.

ZHANG Y, GALLEY M, GAO J, et al. Generating informative and diverse conversational responses via adversarial information maximization[C/OL]//BENGIO S, WALLACH H M, LAROCHELLE H, et al. Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. 2018c: 1815-1825. https://proceedings.neurips.cc/paper/2018/hash/23ce1851341ec1fa9e0c259de10bf87c-Abstract.html.

ZHANG Y, GAO X, LEE S, et al. Consistent dialogue generation with self-supervised feature learning[J]. ArXiv, 2019, abs/1903.05759.

ZHANG Y, SUN S, GALLEY M, et al. DIALOGPT : Large-scale generative pre-training for conversational response generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Online, 2020a: 270-278.

ZHANG Y, SUN S, GALLEY M, et al. DIALOGPT : Large-scale generative pre-training for conver-

sational response generation[C/OL]//ÇELIKYILMAZ A, WEN T. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, 2020b: 270-278. https://doi.org/10.18653/v1/2020.acl-demos.30.

ZHAO X, WU W, XU C, et al. Knowledge-grounded dialogue generation with pre-trained language models[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020: 3377-3390.

ZHENG Y, CHEN G, HUANG M, et al. Personalized dialogue generation with diversified traits[J]. 2019.

ZHOU H, YOUNG T, HUANG M, et al. Commonsense knowledge aware conversation generation with graph attention[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. International Joint Conferences on Artificial Intelligence Organization, 2018a: 4623-4629.

ZHOU K, PRABHUMOYE S, BLACK A W. A dataset for document grounded conversations [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. 2018b: 708-713.

ZHOU X, DONG D, WU H, et al. Multi-view response selection for human-computer conversation [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016: 372-381.

ZHOU X, LI L, DONG D, et al. Multi-turn response selection for chatbots with deep attention matching network[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018c: 1118-1127.

致谢

时光荏苒，三年的硕士研究生涯马上就要结束。在此毕业之际，我要特别感谢所有帮助和支持过我的人，包括一直以来全力支持我的家人们、悉心指导我的老师们、无私帮助我的师兄师姐师弟师妹们、以及陪伴我走过这三年时光的同学们。

首先我要由衷地感谢我的研究生导师冯洋老师。依稀记得四年前，冯洋老师录取我进入研究组，让我有机会开启一段受益匪浅的求学经历。冯洋老师渊博深厚的学识、勤奋认真的科研态度无时无刻不在激励着我，还记得我凌晨一点给冯洋老师发消息她都能秒回。无论是科研受挫还是生活不顺，在我遇到困难时她总是能伸出援手并给予充分的理解。感谢冯洋老师给我提供了优越舒适的科研环境、全方位的指导和极大的信任，让我可以在学术道路上自由地探索，让我学会从整体思考研究方向和研究内容，让我学会如何做科研，更学会如何做有价值的科研。

感谢我在腾讯微信模式识别中心三年校企合作期间所有同事对我的帮助。尤其感谢张金超师兄和孟凡东师兄。张金超师兄博览群书、风趣幽默，对研究有极好的品味和逻辑，教会我如何去做有价值的研究。孟凡东师兄是我科研路上的引路人，给予了我很多学术研究和职业规划上的指导。感谢牛成博士、周杰博士，为我提供了良好的科研环境和学术指导，也在职业发展上给予了我非常有价值的建议。三年的校企合作经历让我在之后的工作中受益匪浅。

感谢实验室的师兄师姐们对我生活和学术上的帮助。感谢杨郑鑫、谷舒豪、邵晨泽、单勇、王树根师兄和刘舒曼、李京谕、申磊、欧蛟师姐，师兄师姐们在科研道路上授予了我宝贵的经验和帮助，令我受益颇丰。感谢谷舒豪师兄和单勇师兄，感谢谷舒豪师兄为实验室传承和团队建设做出的努力，单勇师兄是我本科和研究生的直系师兄，感谢单勇师兄在我保研时向冯洋老师极力推荐，我才有机会进入到研究组。感谢和我同级的郭登级同学，有幸和你共同度过三年快乐的学习时光，一起组织春游，一起组织年终晚会的经历还历历在目。感谢研究组的师弟师妹们：张绍磊、伍烜甫、张倬诚、马铮睿、房庆凯、刘龙祥、黄浪林、桂绍彤、赵彤钰、郭守涛、杨哲，你们是研究组发展的中坚力量和未来的希望。

感谢 15015 的室友，姚青松、李珺、全权、冯天海，很开心与你们一起生活，一起聚餐，一起开黑。希望 15015 的论文、荣誉越来越多。聚是一团火，散作满天星。

感谢华中科技大学 Dian 团队对我的培养和支持，让我在毕业后仍时常能感受到来自团队的关怀、回忆起本科时与 “战友” 并肩作战的青葱岁月。感谢刘玉老师、王兴刚老师在职业发展上给予我的非常有价值的建议，感谢黄涛、陆华、杨澍生、杨阳、孙昊海、钟嘉伦等师弟师妹们在我需要时提供的帮助与合作。感谢符史梁、王淇营、许洪深在我生活上遇到挫折时给我提供的帮助。

特别感谢我的父母、妹妹和其他家人，你们是我无论身在何处都能依靠的温馨港湾、是我砥砺前行的精神支柱，衷心祝愿你们身体健康，幸福平安。感谢我的亲人们对我一直以来的包容和照顾。

最后感谢在百忙之中评阅论文并提出宝贵意见的各位老师，以及在此感谢所有帮助过我的人。

作者简历及攻读学位期间发表的学术论文与研究成果

基本情况

姓名：李泽康　性别：男　出生日期：1997.09.10　籍贯：河北省石家庄市

教育经历

2015 年 09 月–2019 年 06 月，于华中科技大学电子信息与通信学院获得学士学位。

2019 年 09 月–2022 年 06 月，于中国科学院计算技术研究所攻读硕士学位。

攻读硕士学位期间的获奖情况

2019.10　国际对话技术竞赛 DSTC8，多模态对话赛道，冠军

2020.10　国际对话技术竞赛 DSTC9，交互式对话赛道，任务一冠军，任务二第三名

2020.10　中国科学院计算技术研究所 “学业奖学金” 一等奖。

2020.12　中国科学院计算技术研究所 “虑得投资硕士生奖”。

2021.04　中国科学院大学 “三好学生” 荣誉称号。

2021.10　中国科学院计算技术研究所 “学业奖学金” 一等奖。

2021.12　中国科学院大学 “国家奖学金”。

攻读硕士学位期间参加的科研项目

1. 国家重点研发计划政府间国际科技创新合作重点专项，项目名称：基于神经网络的汉泰机器翻译研究，项目批准号 2017YFE0192900，2019/08-2022/07

2. 国家重点研发计划科技创新 2030-“新一代人工智能”重大项目子课题，课题名称: 人机行为与情境常识的大规模知识处理与推理，课题号 2018AAA0102502，2019/12-2023/12

3. 国家重点研发计划”前沿科技创新”子课题，子课题名称：面向多语言文本的知识抽取，课题号 2019QY2301，2019/10/31-2021/10/31

攻读硕士学位期间发表的文章

1. **Zekang Li**, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, Jie Zhou, Incremental Transformer with Deliberation Decoder for Document Grounded Conversations, the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), July, 2019, 12–21, Florence, Italy, 2019

2. **Zekang Li**, Zongjia Li, Jinchao Zhang, Yang Feng, Jie Zhou. Bridging Text and Video: A Universal Multimodal Transformer for Video-Audio Scene-Aware Dialog. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 29, 2476-2483

3. **Zekang Li**, Jinchao Zhang, Zhengcong Fei, Yang Feng, Jie Zhou. Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances. The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), August 1-6, 2021, 128–138, online

4. **Zekang Li**, Jinchao Zhang, Zhengcong Fei, Yang Feng, Jie Zhou. Addressing Inquiries about History: An Efficient and Practical Framework for Evaluating Open-domain Chatbot Consistency. The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Findings of ACL-IJCNLP 2021), August 1-6, 2021, 1057–1067, online

5. **Zekang Li**, Zongjia Li, Jinchao Zhang, Yang Feng, Jie Zhou. WeChat AI & ICT's Submission for DSTC9 Interactive Dialogue Evaluation Track. Proceedings of the 9th Dialog System Technology Challenge Workshop in AAAI2021.

6. Tao Huang, **Zekang Li**, Hua Lu, Yong Shan, Shusheng Yang, Yang Feng, Fei Wang, Shan You, Chang Xu. Relational Surrogate Loss Learning. International Conference on Learning Representations (ICLR), 2022.

7. Zhengcong Fei, **Zekang Li**, Jinchao Zhang, Yang Feng, Jie Zhou. Towards Expressive Communication with Internet Memes: A New Multimodal Conversation Dataset and Benchmark. Proceedings of the 10th Dialog System Technology Challenge Workshop in AAAI2022.

联系方式

通讯地址：北京市海淀区科学院南路 6 号中国科学院计算技术研究所

邮编：100190

E-mail:lizekang19g@ict.ac.cn