

本科生毕业论文

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 个性化生成式开放域对话系统的研究 | | |
| 学 院 | 计算机与信息工程学院 | | |
| 学生姓名 | 赵亭亭 | 学号 | 2019122106476 |
| 专 业 | 计算机科学与技术 | 年级 | 19级 |
| 指导教师 | 王冬青 | 职称 | 讲师 |

内蒙古农业大学教务处制

2023年 5 月

摘 要

随着人工智能和自然语言处理技术的不断进步和发展，开放域（闲聊型）对话系统也在不断发展。开放域对话系统可以与用户进行无限制对话，不会受到话题或情境的限制，可以在广泛的话题范围内进行对话并持续学习。

为了保证对话回复的连贯性，本系统建立了全局唯一的虚拟人设，将用户的对话历史记录、对话主题等信息进行管理和维护。在对话过程中，系统将用户的对话历史记录和当前的对话主题进行分析和挖掘，以便更好地理解用户的需求和意图，并能够更加准确地生成回复内容，从而增强了系统的交互性和真实感，实现了高质量的自然语言交互。

在具体实现方面，本系统采用了CPM大模型，以及Transformer架构，能够在大规模数据集上进行训练和预测，并且具有强大的自适应学习能力。在训练过程中，使用了PyTorch库，能够更加高效和方便地实现对话模型的训练和优化。

关键词：开放域对话系统；生成式；个性化；神经网络；Transformer

Research on Personalized Generative and Retrieval Dialogue Systems Based on Python

Abstract

With the continuous progress and development of artificial intelligence and natural language processing technology, open domain (chat) dialogue systems are also developing. The open domain dialogue system can engage in unrestricted conversations with users, without being limited by topics or contexts, and can engage in conversations and continue learning across a wide range of topics.

In order to ensure the coherence of conversation responses, this system has established a globally unique virtual human setting, which manages and maintains users' conversation history, conversation topics, and other information. During the dialogue process, the system analyzes and mines the user's conversation history and current conversation topics to better understand their needs and intentions, and can generate reply content more accurately, thereby enhancing the system's interactivity and realism, and achieving high-quality natural language interaction.

In terms of specific implementation, this system adopts the CPM large model and Transformer architecture, which can train and predict on large-scale datasets and has strong adaptive learning ability. During the training process, the PyTorch library was used to achieve more efficient and convenient training and optimization of the dialogue model.

Key Words: *Open domain dialogue system; Generative; Individualization;Neural network;Transformer*

目 录

[1 绪论 1](#_Toc30532)

[1.1研究背景及意义 1](#_Toc29188)

[1.2国内外研究现状 1](#_Toc21678)

[1.3主要研究内容 2](#_Toc20893)

[1.4 论文组织框架 2](#_Toc21241)

[2 相关技术介绍 3](#_Toc29082)

[2.1.开放域对话系统概述 3](#_Toc19465)

[2.1.1基于生成式技术的对话系统 3](#_Toc32375)

[2.1.2基于端到端技术的对话系统 3](#_Toc13291)

[2.2 自然语言理解与生成技术 4](#_Toc14216)

[2.3大模型 4](#_Toc25139)

[2.3.1 transformer 4](#_Toc7512)

[2.3.2 PyTorch框架 5](#_Toc26471)

[2.3.3 CPM 6](#_Toc20470)

[2.4意图识别算法 6](#_Toc22153)

[3 开放域对话系统的设计与实现 7](#_Toc1214)

[3.1 需求分析 7](#_Toc27719)

[3.1.1存在的问题 7](#_Toc30562)

[3.1.2本系统功能需求 7](#_Toc2140)

[3.2系统的设计与实现 8](#_Toc6206)

[3.2.1系统整体架构 8](#_Toc21273)

[3.2.2数据来源以及输入形式 9](#_Toc30780)

[3.3数据构造与解析 10](#_Toc21693)

[3.4模型训练 11](#_Toc27005)

[3.5实验测试 12](#_Toc4770)

[4 总结与展望 14](#_Toc26785)

[4.1总结 14](#_Toc27355)

[4.2展望 14](#_Toc2509)

[致 谢 15](#_Toc6)

[参 考 文 献 16](#_Toc6237)

# 1 绪论

## 1.1研究背景及意义

对话系统是人工智能技术的综合应用,涉及到自然语言理解、自然语言生成、对话模型、知识表示及推理、机器学习等诸多方面,是机器智能的一种综合体现[1]。对话模型的研究起源于图灵测试，图灵测试在1950年由Alan Turing提出,其目的是用人-机对话方式机械地验证机器是否能够具有智能,进行多次测试后，如果机器让平均每个参与者做出超过30%的误判，那么这台机器就通过了测试，并被认为具有智能[2]。图灵测试从本质上讲是纯粹的语言行为对智能的体现，人-机对话是其表现形式。2014年，在英国皇家会举行的“2014图灵测试”大会上，聊天程序“尤金古特斯曼”（Eugene Goostman）首次通过了图灵测试，至此，对话系统掀起了新的浪潮[3]。

开放域对话系统技术有着非常重要的价值，是人工智能发展的一个重要分支，是当前众多科技公司的研究热点，在各种应用上为人们提供服务。但目前大部分的生成式与检索式对话系统仍然存在一些问题，比如产生无意义对话，上下文矛盾、无法完成连续的多轮对话等。针对以上问题，研究通过文献研究法与案例研究法，提高对话内容质量。

## 1.2国内外研究现状

自图灵测试提出以来，人工智能对话一直备受国际计算机学界的高度关注和重视，成为诸多学者关注的热点。随着现代计算机计算能力的显著提升和互联网的广泛普及，产生了海量的数据资源。深度学习等新一代人工智能技术的崛起，自然语言处理技术得到了广泛的研究和应用。其中，人机对话系统作为自然语言处理领域最受关注的应用之一，已经成为了研究的重点。在智能对话研究领域中，关于开放域对话系统方向的研究，国内外学者取得了不错的成果。

基于规则或基于模板的对话系统是第一代技术；随着大数据技术的发展，统计机器翻译在自动翻译上取得了巨大成功，热潮蔓延到对话系统领域，出现了统计对话系统，为第二代技术；2014年，深度学习在计算机视觉和语音识别上的突破性进展，以及深度增强学习在Atari游戏上的成功，使学者们开始将诸多神经网络模型用于对话系统，称为神经对话系统，为第三代技术[4]。

实现人机语音对话，可以彻底改善人机界面，从而使更多的人能更方便的与计算机实现交流。由于人机对话言语工程技术的重要性，我国一直在致力于人机对话系统的研究，在“八五”期间，由中科院声学所、自动化所和中国科技大学所承担的中科院重大科研项目“汉语人-机语音对话系统工程”取得了重大科研成果[5]。该项目完成的“北京旅游信息咨询系统”是第一个比较完整的汉语人机对话系统[6]。2021年中国移动九天人工智能新技术论坛上提出，要建设强泛化人机语音对话系统迈向强人工智能。腾讯云智能通过整合智能对话系统与知识图谱系统中的AI基础能力，提出的多语言预测训练模型“神农MShenNonG”[7]，多次登顶业界重要榜单。

一些发达国家如美国、日本和西欧各国也都在这个领域有庞大的研究计划和大量的资金投入。AT&T的语音拨号系统、MIT和CPU的航空公司自动订票系统等都是人机语音对话系统在特定领域的应用系统过渡的阶段[8]。

2011年IBM 的WOTSON在美国电视知识抢答节目上的表现优秀，再次把人们的注意力聚焦在人机对话上[9]；同年，苹果手机SIRI发布掀起了手机语音对话的序幕；2014年亚马逊的ECHO发布以来，持续的大量投入，从2017年开始至今，每年举行Alexa Priza Challenges国际竞赛，奖金高达50万美金，竞赛的火爆也说明了人机对话研究的关注程度；而微软也不落人后，紧随其后退出了社交机器人小冰[10]。

2022年11月30日，美国OpenAI研发的聊天机器人程序ChatGPT正式发布（Chat Generative Pre-trained Transformer），ChatGPT能够通过理解和学习人类的语言来进行对话，还能根据聊天的上下文进行互动，真正像人类一样来聊天交流，甚至能完成撰写邮件、视频脚本、文案、翻译、代码，写论文等任务，一经发布，这款高智能的对话式人工智能便在全球范围狂揽1亿名用户，引发众多科技公司的追逐[11]。

## 1.3主要研究内容

对智能对话系统的现状进行分析，整理国内外研究成果，基于已有方法探讨改进以下问题：

1. 改善上下文矛盾问题，使对话更加连贯；
2. 使开放域对话系统更具个性化，具备稳定的人格特征,有完善的个人历史,能够和现实世界的一些内容交互（例如天气、新闻等），具有长期角色信息能力；
3. 使智能对话的回复更加有意义，能实现连续的多轮对话。

## 1.4 论文组织框架

第一章是绪论部分，综合叙述本研究的背景与意义，分析了目前的研究现状，结合目前研究中所存在的问题，提出主要研究目标与内容。

第二章介绍论文相关知识背景和涉及到的算法。首先，介绍了对话系统的基本分类，包括任务型对话系统和闲聊型对话系统，然后介绍了自然语言理解与生成技术，最后，介绍本文涉及到的大模型与意识识别算法。

第三章介绍系统设计部分，介绍了闲聊型对话的设计与实现，结合前一章提出的算法，对闲聊型对话系统进行了需求分析、模块设计，并进行开发与测试。

第四章对研究工作进行了总结，对未来的研究方向进行了展望。

# 2 相关技术介绍

## 2.1.开放域对话系统概述

### 2.1.1基于生成式技术的对话系统

论文研究的开放域对话系统是基于生成式技术的聊天对话系统。生成式对话系统是一种常见的自然语言处理应用，它能够根据对话历史和当前对话情境，动态生成符合语法和语义规则的自然语言文本，以实现多轮对话交互。与检索式对话系统相比，生成式对话系统更加灵活和多样化，能够自动生成符合语义和语法规则的文本，并能够结合上下文信息和语境进行回复，实现更加自然和流畅的对话交互。

在具体实现中，生成式对话系统一般采用序列到序列的生成模型，如基于Transformer架构的GPT系列模型。在训练过程中，系统通过大规模的语料库学习语言模型，并利用编码器-解码器框架生成合适的回复文本。而在测试过程中，系统将对话历史序列作为输入，通过解码器生成回复文本，并根据上下文信息、用户意图等因素生成合适的回复文本。随着模型的不断训练和调整，生成式对话系统能够学习到更加复杂和抽象的语言能力，实现高质量、个性化的回复。

然而，生成式对话系统的缺点在于，生成的回复文本可能会存在流畅性不佳、内容的连贯性和一致性等问题，也可能存在生成不符合对话语境的回复的情况。为了克服这些问题，需要继续深入研究和优化，加强对话历史的管理和学习，提升对话质量和用户满意度。

### 2.1.2基于端到端技术的对话系统

端到端模型建模已经逐渐成为开放域对话任务的主流方案。基于端到端技术的对话系统是指在一个框架内完成自然语言理解和自然语言生成的对话系统，它通过整合多个技术组件，如语义理解、对话管理、语音识别、自然语言生成等，将对话系统的不同技术组件合并为一个单一的系统，从而达到更加高效、连续和自然的对话体验。

常见的端到端对话系统架构包括：基于循环神经网络 (RNN) 的模型[12]、基于转移学习的模型、注意力机制模型等。 在这些模型中，循环神经网络模型经常被用于处理长序列的文本，而注意力机制模型被用于在不同的输入和输出时刻进行信息的自适应分配和加权，从而提高了系统的学习和预测性能。

端到端对话系统的优势包括：

1. 自适应性强：端到端对话系统可以根据上下文和用户输入自适应地推断和生成回复，这使得对话系统能够更好地应对用户的需求和歧义。
2. 更好的连续性：端到端对话系统能够实时地处理并回应用户查询或指令，从而实现连续的、真实的对话交流体验。
3. 更强的可拓展性：端到端对话系统可以更好地处理各种自然语言，包括不同的方言、口音和语法规则，实现多语言的支持和应用。
4. 更高的准确性：端到端对话系统通过整合多种技术组件和联合训练来加强对话依存关系和语言规则的学习和理解，从而能够更准确地生成和预测回复。

总的来说，端到端对话系统具有高效、连续、自适应等优势，能够为用户提供更加智能和自然的对话体验。然而，端到端对话系统的训练和调优需要大量的数据和有效的算法，需要在实际应用中加以进一步改进和提高。

## 2.2 自然语言理解与生成技术

自然语言理解和生成技术是自然语言处理领域中的两个核心技术，它们分别从不同的角度对自然语言进行处理和分析，有着广泛的应用和重要的实际意义。

自然语言理解技术是指将自然语言转换为计算机可读的结构化数据的技术。该技术的目的是将自然语言文本或语音输入解析为计算机可以理解的形式，从中提取有用的信息。常用的自然语言理解技术包括分词、词性标注、句法分析、语义分析、情感分析、实体识别、关系抽取等。

自然语言生成技术是指将结构化数据转换为自然语言的可读的文本的技术。该技术的目的是将结构化数据转换为自然语言，以便计算机与用户进行自然语言的交互。常用的自然语言生成技术包括自然语言模板填充、对话历史记录和上下文管理、文本修正和客户体验优化等。

自然语言理解和生成技术的应用领域广泛，包括智能客服、知识库问答、智能翻译、智能文书写作等多个领域。在智能客服领域中，自然语言理解技术可以帮助系统理解用户的问题和需求，而自然语言生成技术可以帮助系统向用户提供有用的、内容丰富的回复。在知识库问答领域中，自然语言理解技术可以帮助系统理解用户的提问，而自然语言生成技术则可以帮助系统以自然语言形式提供准确、全面和易于理解的答案。因此，自然语言理解和生成技术是自然语言处理技术中非常重要和核心的技术。

## 2.3大模型

### 2.3.1 transformer

论文研究的开放域对话系统主要使用了Transformer模型实现对话生成任务。Transformer是一种利用自注意力机制实现的深度神经网络模型。，本质上是一个从编码器（encoders）到解码器(decoders)的结构，编码器中每个模块由自注意力机制及前馈神经网络组成，解码器的每个模块由自注意力机制、编解码器注意力机制以及前馈神经网络组成[13]。与所有的文本生成模型有一个共同点，编码器的输出都会作为解码器的输入。编码器的任务是将输入序列转化为模型可以处理的向量表示。在编码器中，自注意力机制用于计算输入序列中每个词和其他所有词之间的注意力权重，然后根据这些权重对每个词进行编码。编码器的输出被用于送入解码器，或者直接用于下游任务。解码器的任务是生成目标序列。在解码器中，自注意力机制用于计算目标序列中每一个词和其他所有词之间的注意力权重，序列注意力机制用于计算输入序列和目标序列之间的注意力权重，从而将正确的上下文信息引入生成过程中。解码器每次预测一个单词，并将其作为下一个单词的上下文输入，如图1所示。

Transformer 模型的优点是不需要使用递归神经网络（RNN），相比于 RNN 可以并行计算，更加高效，可以更好地处理长序列和上下文信息。Transformer 模型在机器翻译、文本摘要、问答系统、语音识别等任务中表现出了良好的性能和效果。

近年来，更多的研究工作已经在其基础上发展，如BERT( Bidirectional Encoder Representations from Transformers)。BERT是一种能够双向预训练的Transformer模型，可用于大量自然语言处理任务，极大地推动了自然语言处理的发展[14]。BERT主要用了Transformer的Encoder，而没有用其Decoder，如图2所示，因为BERT是一个预训练模型，只要学到其中语义关系即可，不需要去解码完成具体的任务。transformer 以及其相关变体仍然是自然语言处理领域研究的热点之一。

### 2.3.2 PyTorch框架

论文研究的开放域对话系统使用的深度学习框架是 PyTorch。PyTorch是 Facebook 推出的人工智能学习系统，虽然底层优化是在 C 的基础上实现，但基本所有框架都是用Python 编写，因此源码看起来较简洁明了[15] 。相对于其他框架，它具有创建动态神经网络、支持 GPU、Python 优先等特点。PyTorch的设计追求最小的封装，并遵循三个级别的从低到高的抽象：张量-变量（auto grad）-NM.Module，模块分别代表高维数组（张量），自动推导（变量） 和神经网络（层／模块），这三个抽象密切相关，可以同时修改和操作，相对于 Tensor flow，PyTorch 更加简洁直观[16]。

在PyTorch中，可以用torch.Ann 实现各种神经网络模型，例如卷积神经网络（CNN）、循环神经网络（RNN）、Transformer等。PyTorch还提供了各种优化算法，如随机梯度下降（SGD）、自适应优化算法（Adam、Ada grad）等，来帮助用户训练模型。此外，PyTorch 还提供了专门针对图像、文本等数据类型的预处理库，以及可视化和调试工具，方便用户使用。

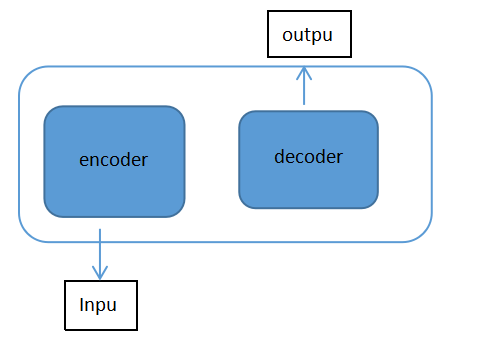


图1 encoder—decoder的Transformer模型

Fig.1 Encoder Decoder Transformer Model

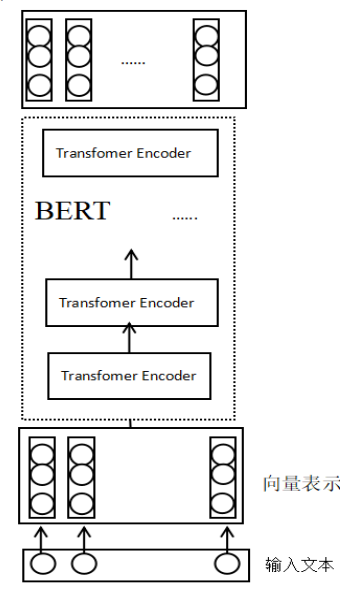


图2 BERT 模型

Fig.2 BERT Model

### 2.3.3 CPM

论文研究的开放域对话系统是在CPM大模型的基础上进行研究。CMP（Chinese Multi-Party dialogues Platform）大模型是由中国科学技术大学自然语言处理与社会人文计算实验室（NLPIR）开发的一款针对中文多轮对话任务的大型预训练语言模型。它是构建在GPT（Generative Pre-trained Transformer）架构的基础上，经过在大规模中文多轮对话数据集上的预训练得到的。 CMP大模型包含了多个模型，分别涵盖不同的模型大小，从3亿到22亿不等，可覆盖不同场景和各种硬件资源的使用需求，从而满足不同应用环境中的需要。该模型在2021年底被提供到了Ali NLP的开放平台上，使得更多的开发者和研究者可以使用该模型进行各种自然语言处理任务。 CMP大模型在多轮对话中具有非常好的表现，因为它能自动学习和理解多轮对话的语义和意图，从而更好地理解对话中的上下文和语境，有效提高了对话的质量和连贯性。该模型可以被广泛应用于各种领域中，包括智能客服、聊天机器人、知识库问答系统等。它的出现在中文自然语言处理领域中有着重要的意义，并为中文自然语言处理技术的发展注入了新的动力。

## 2.4意图识别算法

论文研究的开放域对话系系统通过意图识别算法，研究用户的意图。意图识别（Intent Recognition）是指从自然语言文本或语音中识别用户意图的技术，本质上就是对意图进行分类。通常应用于对话系统、智能客服和语音助手等领域。目前，意图识别技术主要有以下两种算法：

基于规则的意图识别算法，该算法通过手动定义一些规则，如文本中的关键词、短语或特定模式等，用于识别用户的意图。相较于深度学习模型，该算法较为简单，且不需要大量的数据集。然而，该方法需要人工定义规则，因此泛化能力有限，对不同领域、语言和应用情况下的意图识别任务具有一定的局限性。

基于机器学习的意图识别算法，该算法需要根据大量已标注数据来训练一个分类器，通过对输入文本进行特征提取和分类操作来识别用户的意图。Zhao等人提出了一种使用了注意力机制的 CNN 模型，其主要思想是在卷积操作之后引入注意力机制[17]；Yang等人则提出了一种基于层级结构的注意力机制，用于增强模型抓取不同抽象层次概念的能力[18]；Chen等人则提出了使用 BERT 模型进行文本分类的算法，BERT 是一种基于 Transformer 的双向神经网络模型[19]。相较于基于规则的算法，基于机器学习的算法可以自动生成规则，具有较好的泛化能力，能够更好地适应多义词、错别字等语言现象。

值得注意的是，因为意图识别算法在应用中受到许多不确定因素的影响（如社交网络中的缩略语、恶意输入等），因此实际应用需要根据情况进行优化和调整，以满足各种不同的需求和场景。

# 3 开放域对话系统的设计与实现

## 3.1 需求分析

### 3.1.1存在的问题

开放域对话系统能够提供广泛的服务和价值，包括问答、搜索引擎和聊天机器人等，以满足用户的不同需求和目的，从而提供多样化的服务和价值。相较于面向特定领域的对话系统，开放域对话系统面临着更多的挑战和问题，通过对几款比较流行的开放域对话系统包括“AI小呆爱聊天”，“百度PLAOT”,“Emohaa”进行的测试，发现均存在以下问题：

1. 上下文矛盾。三款对话系统都出现了上下文矛盾的问题，系统无法根据对话上下文正确地理解用户的发言，导致回答出现矛盾或非常规想法，与用户的意图和对话的情境不符。

2.对话难以持续。与百度PLATO进行对话，系统无法保持与用户的有效互动和连续对话，无法实现长时间交互。

3.被动回答。当用户向AI小呆爱聊天提出问题时，对话系统给出单一且无足轻重的回答，而不进行对话和互动。

4.无法做出正确响应。测试Emohaa对话系统，发现系统无法根据用户请求和上下文生成准确的回答。这可能会导致用户感到困惑，甚至导致对话中断。

具体测试结果如表1所示，图3为实际测试内容的部分截图。

表1 主流对话系统问题对比

Table.1 Comparison of mainstream dialogue system issues

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 上下文矛盾 | 对话难以持续 | 被动回答 | 无法做出正确响应 |
| AI小呆爱聊天 | √ |  | √ |  |
| 百度PLATO | √ | √ |  |  |
| Emohaa | √ |  |  | √ |



a 上下文矛盾 b 对话难以持续 c 被动回答图 d 无法做出正确响应

图3 三款对话系统的实际测试截图

Fig.3 Actual test screenshots of three dialogue systems

### 3.1.2本系统功能需求

开放域对话系统旨在向用户提供跨领域智能助手能力。通过自然语言的方式与用户进行交流，帮助用户完成具体任务，提供更加高效、个性化、智能的体验。应该满足如下需求目标：

1. 对话系统应该具有固定人设，确保人格稳定。将对话系统设计为具有一定人格和个性特征的实体，可以确保其在多个用户和场景中表现出一致的行为和响应，提高对话系统的稳定性和可靠性，使用户在使用对话系统时更具信心和满意度。
2. 具有意图识别能力。对话系统可以根据用户的提问、答复、所处环境等信息推断出用户的意图，并进行有目的、有针对性的回答，能够感知当前进行的话题，并在合适的时机引导话题或切换到新的话题，进一步提高对话系统的交互性和用户体验。
3. 不出现上下文矛盾。在不把天聊死的情况下对话尽量多的轮数，使对话更加充实丰富，更加符合用户的期望和需求，但又不会出现上下文矛盾或是毫无意义的对话。

## 3.2系统的设计与实现

### 3.2.1系统整体架构

本系统使用Transformer预训练模型处理对话数据，其中BERT是最重要的语料训练模型，BerthaModel是最核心部分，BertLayer是Bert预训练模型中产生句向量和词向量的核心模块，调用关系如图4；Transformer预训练模型在本系统中的重点有以下几点：

（1）BertModel用于对输入的文本进行特征提取，并利用反向传播算法优化模型参数，以提高模型的性能。它首先在forward函数中定义各个参数，然后按键值对的形式向前传播，返回encoder\_output，并通过BertEmbeddings求和得到输入（input）。

（2）BertEncoder进入BertLayer：BertEncoder是Bert模型中的一个模块，由多个BertLayer组成。在对话系统中，BertEncoder可以通过BertLayer和双向Transformer的机制获取上下文信息。BertLayer包括多个子模块，例如BertAttention和FFN。

（3）BertAttention包括BertSelfAttention与BertSelfOutput：BertSelfAttention使用多头注意力机制计算句子中各个位置的权重，从而让模型更好地理解句子的语义结构，是模块的核心区域；BertSelfOutput负责将多头注意力机制的输出与输入进行残差连接，然后进行LayerNorm操作。主要使用Attention机制进行计算，如图5所示，包括四个环节：

1. 得到Q（query）、K（key）和V(value)；
2. 计算得到attention\_scores ；
3. 由softmax计算得到attention\_probs；
4. 计算得到context\_layer。

（4）FFN实现矩阵映射：在BertLayer中的另一个子模块是FFN（Feed-Forward Network），它使用了全连接层的机制将向量映射到新的空间中，从而让模型更好地理解句子的结构和含义。

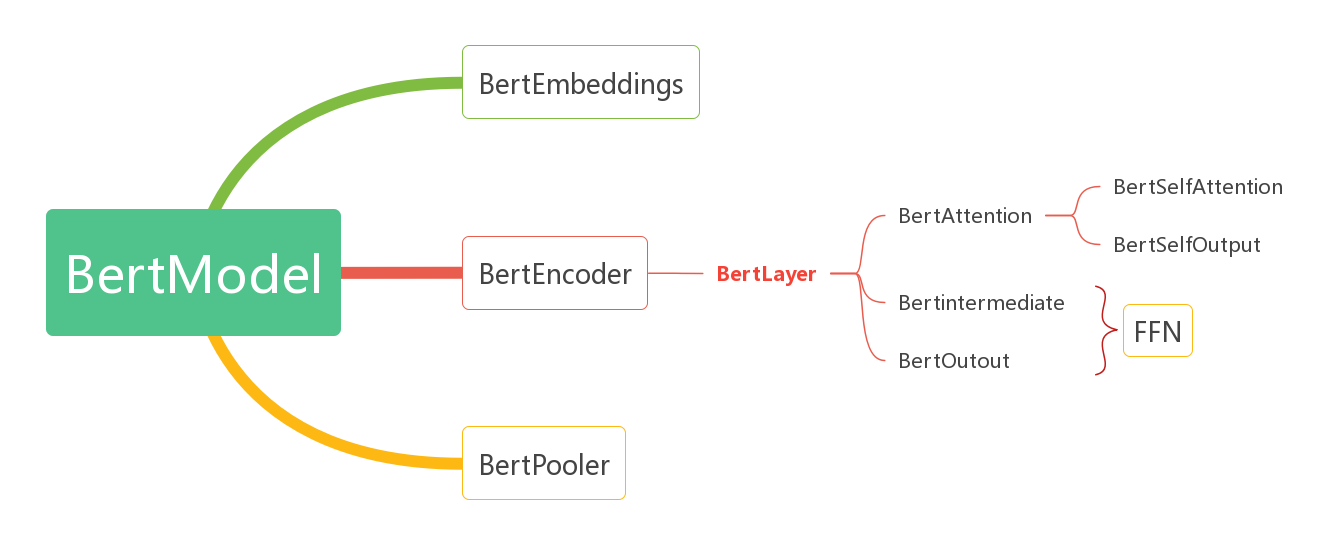


图4 BertModel

Fig.4 BertModel

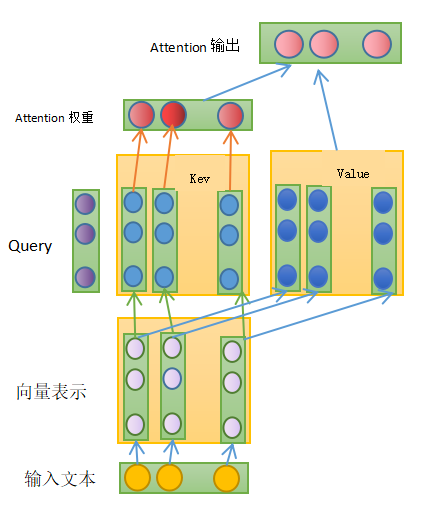


图5 attention机制

Fig.5 Attention mechanism

### 3.2.2数据来源以及输入形式

开放域对话系统系统设计为从Transfomers的encoders（如图6）直接输入（input）数据进行编码，编码以后的结果直接从encoders输出（如图7）。最终将所有的对话数据以字典形式进行统一适配，建立对话系统需要的数据集，为模型的训练和调优提供支持。

对话数据源利用第三方提供的数据集，例如豆瓣、知乎、百度智能等平台提供的API等获取数据，主要通过百度千言数据集，比如DuLeMon（中文长时记忆对话数据集）、Diamante（中文开放域闲聊数据集）、RiSAWOZ（中文多领域任务型对话数据集）、MMChat（中文开放域多模态对话数据集）以及LCCC（开放域短文本对话数据集）等。将数据集下载后包装成.json，再放入模型进行训练，对DuLeMon、Diamante、RiSAWOZ、MMChat以及LCCC对话的平均轮数为16、14、6、2、3轮，最后分析得出他们的对话质量高低，其原因如表2所示。

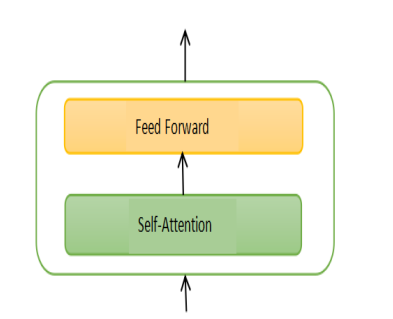
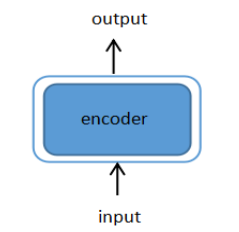
 

图6 encoders的组成 图7 数据输入transformer模型

Fig.6 Composition of encoders Fig.7 Data Input Transformer Model

表2 数据集对比

Table.2 Comparison of Datasets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 数据量 | 平均对话轮数 | 质量高低 | 原因 |
| DuLeMon | 25400 | 16 | 高 | 双方均有个性化 |
| Diamante | 5838 | 14 | 高 | 双人对话，每个对话有候选 |
| RiSAWOZ | 10000 | 6 | 高 | 人机对话，有槽位等信息 |
| MMChat | 19730 | 2 | 中 | 上下文衔接有问题 |
| LCCC | 11987759 | 3 | 低 | 介绍说有上下文知识库，数据中无 |

## 

## 3.3数据构造与解析

对话数据集的构造分为三步，分别是用户画像收集，历史对话收集，拼接对话，其中用户画像收集和历史对话收集部分可以有用户自己设定。

1. 用户画像收集

用户画像主要来自persona数据集，如图8所示，其中对话机器人（part）与用户聊天的角色信息只对自己可见，可以使用其角色信息。用户的角色信息分为两部分：对话机器人已经知道的角色信息、对话机器人不知道的角色信息。这些角色信息用于指导对话过程中的聊天内容和方向。

1. 历史对话收集

针对每个对话，需要根据给定角色信息组织对话，聊天机器人需要利用已知用户角色进行深度聊天，并确保对话内容与给定的信息相关，不能与给定的信息无关，也不能与给定的角色信息相冲突，如图9所示。

1. 将以上两部分进行拼接，即进行对话，如图10所示。通过将用户画像和历史对话进行拼接，可以为聊天机器人提供有价值的对话数据集，帮助聊天机器人更好地理解用户的意图和需求，并更准确地进行回复和推荐。

然后，数据通过递归函数开始解析，如图11所示。

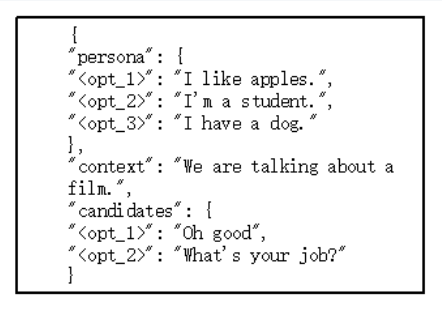
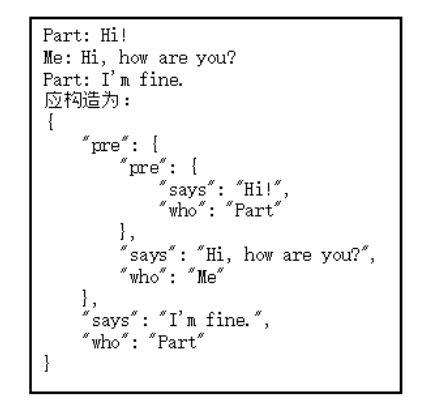
 

图8 用户画像收集 图9 历史对话收集

Fig.8 User Profile Collection Fig.9 Collection of Historical Dialogues

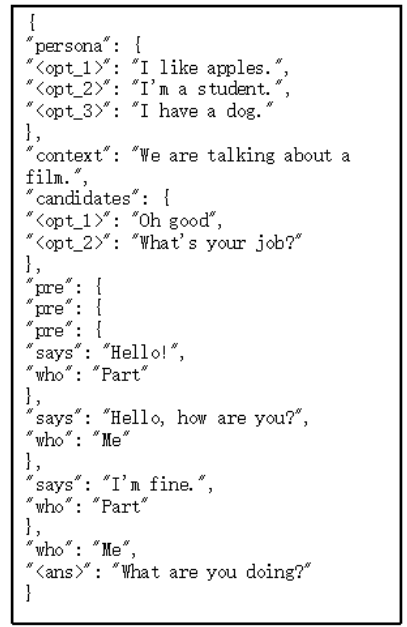


图10 进行对话

Figure 10 Conducting a Dialogue

## 3.4模型训练

对话预训练是一种用于构建自然语言处理应用程序的技术，通过计算损失函数（Loss Function），将真实值与预测值之间的概率分布进行比较，并计算交叉熵，评价模型的预测结果与真实结果之间的误差；再通过自适应学习率的优化器（Adam），自适应的更新学习率，使得模型训练更加高效；最后在对话预训练的模型训练过程中，仔细调整学习率（Learning rate），以确保模型的有效训练。

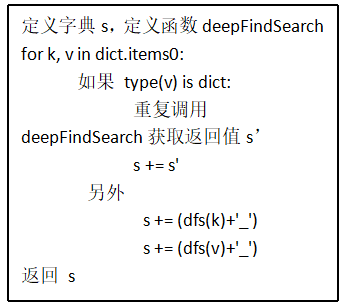


图11 递归函数

Fig.11 Recursive function

## 3.5实验测试

与对话系统进行了50次测试，平均每次测试对话15轮，测试结果如表3，对话系统实现了个性化，如图12所示，能够根据用户的偏好和历史对话记录来定制回复，解决了上下文矛盾的问题；具有了意图识别功能，可以根据用户的提问、答复、所处环境等信息推断出用户的意图，并进行有目的、有针对性的回答；通过循环与字典，将每一次对话输入进行拼接，直到用户输入NO，对话结束，实现了基于上下文进行的多轮连续对话。完整对话测试例子如图13所示。

表3 系统测试

Table.3 System Testing

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试次数 | 平均轮数 | 个性化 | 意图识别 | 多轮连续对话 |
| 50次 | 15轮 | √ | √ | √ |

persona\_list = ['我叫Tom, '我喜欢唱歌 '，我是个男生', '我23岁了', 我喜欢旅游', '我喜欢书法', '我喜欢摇滚乐，我喜欢汪峰', '我是个大学生。']

图12 固定人设

Fig.12 Fixed staffing



图13 完整对话体

Fig.13 Complete dialogue bod

# 4 总结与展望

## 4.1总结

本文对于基于Python的个性化生成式与检索式的对话系统，主要进行了以下研究：

1. 通过personal字段实现了对话系统的个性化，固定了人设，使对话系统具有全局唯一人设；因为对话机器人具有了全局唯一固定人设，因此在对话过程中，机器人不会切换角色，解决了部分上下文矛盾的问题。
2. 通过意图识别算法，将用户的意图挖掘出来，将画像描述出来，并放到训练集里面，从而实现有目的的聊天，使对话系统能够感知当前进行的话题，并在合适的时机引导话题或切换到新话题。
3. 在不出现上下文矛盾，不把天聊死的情况下对话尽量多的轮数。通过字典形式，每一轮对话都可以嵌套在下一轮对话中，逐渐深入话题，并确保对话质量和上下文连贯性，从而实现多轮对话。

对该对话系统进行了测试，测试结果表明对话系统具备个性化能力，可以根据用户的个性化需求进行定制化的回复和推荐；同时，通过聊天上下文的维护，确保了对话的连贯性，使得聊天机器人可以很好地理解用户的意图，并为用户提供更加符合其需求的回复和推荐；另外，在用户未结束对话之前，聊天机器人可以一直与用户进行交互，实现连续的多轮对话。

## 4.2展望

经过测试，对话系统还存在一些可以改进和深入研究的方向，后续可以在以下几个方面继续进行研究：

1. 从互联网中获取知识提升对话质量。可以通过网络爬虫或其他手段，从互联网中获取更加丰富、准确的知识信息，并将其集成到对话系统中，以供回复模板的自动匹配和生成等用途，同时要保证获取的知识信息不会产生矛盾和误导，需要进行适当的过滤和校验。
2. 对历史进行概括，主动拉起新的话题。为了确保对话流程的连贯性和趣味性，可以采用历史对话信息的挖掘和分析方法，提取出对话的关键信息和主题，并以此为基础进行对话推荐和话题扩展，结合智能推荐算法和数据挖掘技术，对用户的兴趣、爱好等信息进行分析和挖掘，并根据其画像信息，提供个性化服务和回复。
3. 控制立场。对于一些具体问题或事件，需要对话系统有明确的立场和观点。可以采用情感分析、舆情监测等技术手段，对当前事件或话题进行分析和评价，并以此为基础，控制对话系统的回复立场和情感倾向，以提高对话的针对性和效果。

参 考 文 献

1. 杨帆。 图灵测试中对话处理研究[J]。 中国科学院数学与系统科学研究所;中国科学院数学与系统科学研究院, 2001。
2. A. M. Turing．Computing Machinery and Intelligence：Oxford University Press on behalf of the Mind Association，1950
3. 雷雨,郭家煊. 通过图灵测试 人工智能就算进入新时代？[N]. 南方日报,2014-07-26(013)
4. 陈龙, & 孙泽健. (2017). 面向任务的对话系统现状研究. 电子技术与软件工程(23),2
5. 张初敏。 人机对话系统工程[C]// "科技增强国力,青年开创未来——携手走向辉煌的新世纪"中国科协第三届青年学术年会。 1998。
6. 向东. 汉语人机语音对话系统工程综述[C]// 中国声学学会青年学术会议. 中国声学学会, 1997.
7. 黄永忠, 张晨昊, & 秦韬. 基于多语言预训练模型的事件触发词检测与分类方法.
8. 赵宇猛. 基于Transformer框架的多轮对话系统模型研究[D].辽宁工程技术大学,2021.DOI:10.27210/d.cnki.glnju.2021.000133.
9. Barton M . When Watson met Siri: Apple's IBM deal could make Siri a lot smarter - Innovation Insights.
10. 孙冰. 微软"小冰"的爆红与暴毙[J]. 中国经济周刊, 2014, 000(022):64-65.
11. 蒲清平,向往.生成式人工智能——ChatGPT的变革影响、风险挑战及应对策略[J/OL].重庆大学学报(社会科学版):1-13[2023-04-14].http://kns.cnki.net/kcms/detail/50.1023.C.20230412.1004.002.html.
12. Goodfellow, I., Bengio, Y., Courville, A.．Deep learning (Vol. 1)：Cambridge：MIT Press，2016：367-415
13. Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.
14. Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
15. Paszke A , Gross S , Massa F , et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library[J]. 2019.
16. 申仲峰. 基于PyTorch框架下北方田地常见杂草的识别[D].山西农业大学,2019.DOI:10.27285/d.cnki.gsxnu.2019.000658.
17. Zhao Z, Wu Y. Attention-Based convolutional neural networks for sentence

classification[C]. INTERSPEECH. 2016: 705-709.

1. Chen Q, Zhuo Z, Wang W. Bert for joint intent classification and slot filling[J]. arXiv preprint arXiv:1902.10909, 2019.
2. Yang Z , Yang D , Dyer C , et al. Hierarchical Attention Networks for Document Classification[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.