自然语言处理课程实践项目报告

面向人机共情的对话生成系统

姓　　名：蔡铭修

学　　号：2201760

专业名称：计算机科学与技术

课程名称：自然语言处理

授课教师：肖桐 教授

二〇二二年秋季

面向人机共情的对话生成系统

1 项目概述

1.1 研究目的

对话系统是人工智能领域的一个非常重要的研究方向，研究者认为使机器具有智能和语言能力是一项极具挑战性的工作。基于生成的对话系统从对话语料中习得对话能力，对用户话语进行直接生成回复，生成的回复效果较好且灵活性较强，适用于开放域对话。近年来，随着深度学习技术的突破性进展以及神经网络的兴起，人工智能领域有了重大的进展。社交网络的不断普及产生了大量的对话数据，为对话系统的研究提供了丰富多样的语料。研究人员开始用深度学习技术进行基于生成的对话的研究，利用神经网络从大量的数据中自动提取特征，对用户输入的对话进行理解，并学习语言能力进行回复话语的生成，大大提升了对话系统的交互效果。

情绪是人类智能的重要组成部分，在信息交流中有着不可替代的作用。人工智能研究的一个重要目标就是让机器能够理解并且表达出情绪。一个能与人类畅谈的机器人，不仅需要理解和表达语言，更重要的是能够理解情绪并将情绪表达融入语言，拥有智商与情商的结合。

随着社会的不断发展，人们对于人机交互体验的期待值也在不断上升，不再满足于呆板的交互，而是期待更加自然、情感化的人机交互体验。然而一直以来，许多对话系统的研究都将着眼点放在对话质量的提升，而对于情绪元素的加入则较少涉及，直到近几年才有关于共情对话的相关研究。显而易见的是，共情对话不是对话系统与情绪分析任务的简单叠加，面对人机对话中不断变化的语义和语境，给出适当的共情回复是比较复杂的，需要动态结合对话内容理解和情绪融入两部分。因此，如何让对话系统更好地理解对话情绪并生成合适的情绪回复，是共情对话系统研究所面临的重要挑战，也是本次课程实践项目研究的核心。

1.2 当前挑战

虽然目前对人机共情对话生成的研究已经有了比较多的研究成果，但是由于

人类的情绪表达是复杂的，因此当前共情对话生成方向仍然面临一些挑战，具体如下：

（1）对于共情对话中用户情绪识别部分，目前较多的研究将用户情绪分为积极、中性和消极三种，其分类粒度不够细化且准确率有待提高。

（2）大部分研究中输入话语与生成话语的情绪一致，而未根据心理学进行回复情绪预测，缺乏合理性。

（3）共情对话生成部分常采用带有情绪标签的对话语料进行有监督训练，这样生成的对话共情程度不足。

（4）生成语句容易出现安全回复，多样化程度较低。

本次课程实践项目针对共情对话生成方向的四个挑战来进行研究，以构建一种能够准确识别用户情绪、确定回复情绪，并能够生成共情且多样化回复的人机共情对话生成系统为最终的目标。

1.3 课程实践项目主要工作

本次课程实践项目主要面向上一节描述的四个挑战进行共情对话系统的研究和实现。根据对话生成过程的通常步骤（对话理解、对话管理、对话生成）及情绪对话的特殊要求（情绪分析、情绪预测），本项目主要研究内容有以下几个方面：

（1）用户话语情绪分析

首先整合了 7 分类的细粒度情绪分类数据集 OCEMOTION 和 NLPCC 2013 中文微博细粒度情感分类数据集作为模型的数据集，并对数据集进行预处理和划分；对 BERT 预训练模型进行下游微调，并采用提示学习技术来构建模板，在合适的位置添加[MASK]符号，同时设置了情绪映射词表。最终训练得到的模型可实现通过情绪分类器分析用户输入话语中的情绪，对情绪进行分类。

（2）回复话语情绪预测

回复话语预测基于情绪知识图谱，因此首先基于情绪心理学知识以及中文情感词汇本体库，构建面向共情关系的情绪知识图谱，接着基于 Neo4j 图数据库实现了情绪知识图谱的存储与可视化，然后基于系统的用户情绪分析结果，通过情绪知识图谱进行 Cypher 查询来预测相应的回复情绪。

（3）共情回复话语生成

首先选取了大规模的中文对话语料，并对语料进行预处理，作为共情回复生成模型的数据集；选取预练模型 GPT-2 进行下游训练，将多轮对话建模为长文本，成为文本生成任务，对于用户输入话语可生成 8 个候选回复；同时在模型中加入互信息，使回复更具有多样性，避免出现诸如“我不知道”这样的安全回复。最后调用情绪分类器对候选回复进行情绪分类，从符合预测回复情绪的候选回复中选取互信息 loss 最小的作为对话回复，生成了具有情绪的回复。

（4）人机共情对话原型系统

人机共情对话原型系统的主要功能是实现人机交互，集前三部分的方法和训练完成的模型等研究成果于一体，采用 Web 技术开发出一套原型系统。共情对话原型系统的呈现形式是一个文本输入、输出形式的人机对话 Web 页面，实现用户与对话系统的自然流畅的交互。

2 用户话语情绪分析

2.1 问题描述

对于用户输入的对话内容来说，情绪信息属于隐藏的信息，模型无法直接进行识别，需要采用文本分析的方法对用户输入话语中的情绪进行提取，得到每句对话的情绪信息，从而进行接下来的回复生成任务。因此，需要对输入话语进行情绪分析的工作，作为整个系统中的情绪分类任务。

总的来说，情绪分类任务可以用以下文本来描述：给定用户输入话语的文本词序列 $A = \{a_1, a_2, \dots, a_x, \dots, a_n\}$，情绪分类模型输出用户话语的情绪标签 e。在这里面文本序列长度为 n，a_x 为文本序列的第 x 个词语，情绪 e 共有七种，分别是悲伤、高兴、难过、愤怒、喜欢、惊奇和恐惧。

2.2 情绪分类模型

本项目选用 BERT 预训练模型作为基础模型，通过对其进行下游的提示微调来构建成一个情绪分类模型。BERT 是一个常用的预训练语言模型，它采用了掩码语言模型（MLM），以此来代替传统单向语言模型以及将两个单向模型浅层次

进行拼接的方法，这使得它拥有了更强的特征抽取方面的能力。

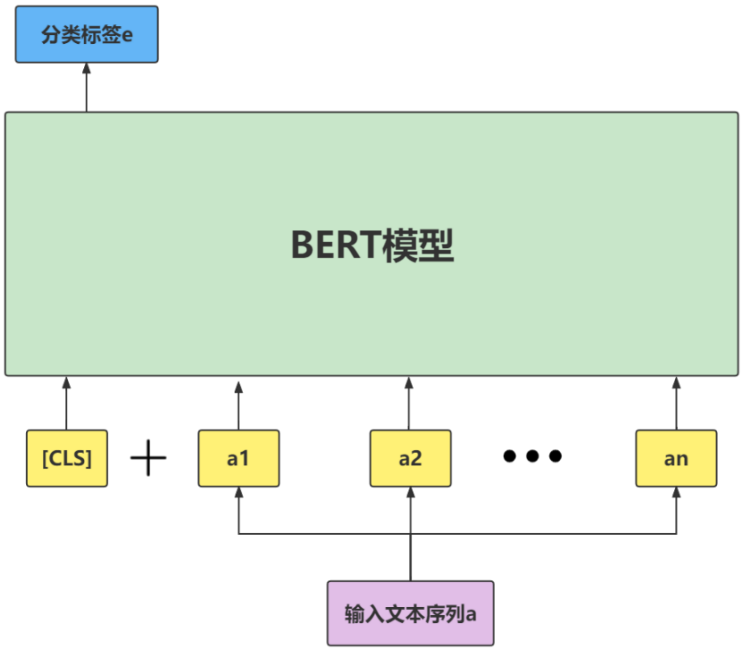

图 2.1 BERT 处理分类任务示意图

BERT 模型首先在大规模语料上进行训练，生成预训练模型，然后针对特定任务进行下游微调。

BERT 模型处理分类任务的过程如图 2.1 所示，可以看到 BERT 模型通过下游微调后进行情绪语句分类任务的整体过程是首先将代表分类的符号[CLS]和文本序列进行拼接，然后输入模型，输入序列依次经过嵌入层和几层 Transformer 编码器之后，最终的结果是输出七分类情绪标签。

同时，为了充分激发 BERT 预训练模型的潜能，利用 BERT 自身的能力来提升分类的准确性，本项目使用提示学习（Prompt learning）的方法对 BERT 模型的微调方式进行改进。基于提示学习的 BERT 模型对输入文本序列的概率进行建模，并使用这一概率来预测分类标签，以此来减少模型的训练对大量数据的需求量，可以在样本量较少的情况下也能得到好的效果。

提示学习充分利用了预训练模型所具有的能力，将情绪分类的任务转换为类似于完形填空的任务。在 BERT 的预训练阶段，其中的一个任务是随机地对一些单词进行 MASK 操作，接着根据样本的上下文预测这些单词的内容，此过程采用了 BERT 的最后一个编码器，然后对其连接一个全连接层，做 Softmax 分类。普通的 BERT 下游微调抛弃了全连接层，用一个新的初始化层进行任务的训练，

这样的话显然抛弃了很多参数，提示学习将这些参数加以利用，针对本项目的用户话语情绪分析任务来说，将情绪类别标签融入模板中，并且用两个[MASK]符号将其替换，通过预训练模型本身的预测能力对 MASK 掉部分进行预测[错误!未找到引用源。]。

本项目基于提示学习的 BERT 模型建立的步骤分为构建模板、构建映射、进行预测。首先，构建提示学习模板，本项目使用的模板构建格式及示例如下：

模板：句子+“，情绪是[MASK][MASK]”

样例：今天考试考砸了，我好伤心

句子+模板：今天考试考砸了，我好伤心，情绪是[MASK][MASK]

由对样例句构建模板的例子可见，本项目通过提示学习将情绪分类的任务转换为类似于完形填空的任务，在此过程中，模板的设计形式显然对于最终模型的效果好坏起着重要的作用。

然后，需要构建映射词表。由于 BERT 预训练模型在进行完形填空任务的过程中，生成文本的范围并没有限制，因此预测生成的词语并不局限于七个情绪标签词：悲伤，高兴，难过，愤怒，喜欢，惊奇，恐惧。由此可见，需要建立一个映射词表，与情绪标签词进行映射。本项目的映射词表构建参考了开源的中文情感词汇本体库，如表 2.1 所示。中文情感词汇本体库的情感分为七类，与本项目的七个情绪标签词一一对应，另外本体库中每种情感对应了大量的例词，这些例词蕴含了对应的情感，本项目选取了例词中的两字词语作为映射词表的素材，完成了分别对应七种情绪的映射词表。经过模板和映射的构建，模型基本建立完成，此时可以进行模型的训练和预测工作。

表 2.1 中文情感词汇本体库分类表

编号	情感大类	情感类	例词
1	乐	快乐	喜悦、欢喜、笑眯眯
2		安心	踏实、宽心、定心丸
3	好	尊敬	恭敬、敬爱、毕恭毕敬
4		赞扬	英俊、优秀、通情达理
5		相信	信任、信赖、可靠
6		喜爱	倾慕、宝贝、一见钟情
7		祝愿	渴望、保佑、福寿绵长

8	怒	愤怒	气愤、恼火、大发雷霆
9	哀	悲伤	忧伤、悲苦、心如刀割
10		失望	憾事、绝望、灰心丧气
11		疚	内疚、忏悔、过意不去
12		思	思念、相思、牵肠挂肚
13	惧	慌	慌张、心慌、不知所措
14		恐惧	胆怯、害怕、担惊受怕
15		羞	害羞、害躁、面红耳赤
16	恶	烦闷	憋闷、烦躁、心烦意乱
17		憎恶	反感、可耻、恨之入骨
18		贬责	呆板、虚荣、杂乱无章
19		嫉妒	眼红、吃醋、醋坛子
20		怀疑	多心、生疑、将信将疑
21	惊	惊奇	奇怪、奇迹、大吃一惊

在基于提示学习的 BERT 预训练模型进行预测时，使用了 torch.topk()函数，取模型生成概率最高的五个词语，若此五个词语中有某一个词语在七个情绪映射表，则分类结果为该对应的感情；若五个词语中的多个分别在不同映射表中，则选取概率最高的词语所在的情绪映射表对应的情绪作为分类结果。

本项目的用户话语情绪分类模型整体架构如图 2.2 所示。

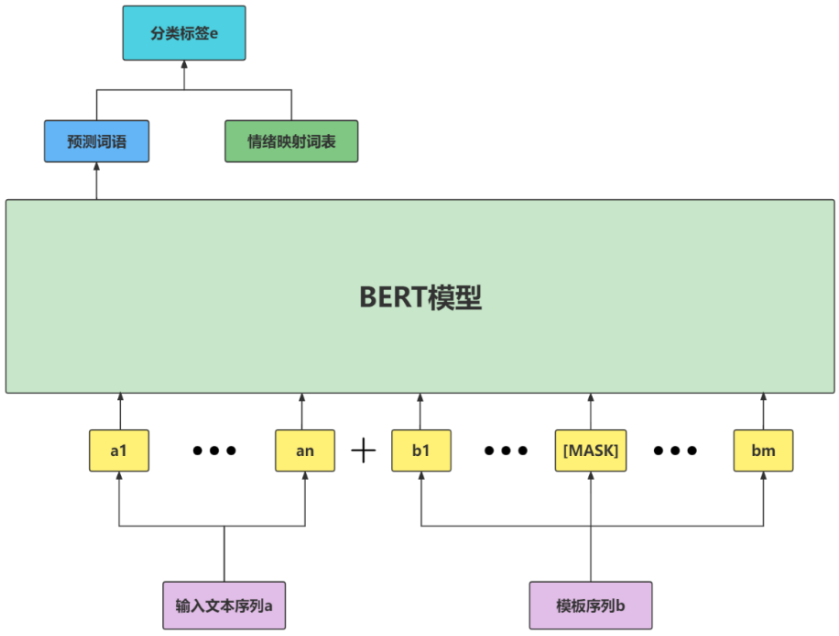

图 2.2　BERT 提示微调进行分类任务示意图

总的来说，用户话语情绪分类模型总体构建和训练的算法流程如表 2.2 所示。

表 2.2 用户话语情绪分类模型算法流程

输入：情绪分类数据集；
输出：用户话语情绪分类模型；
1：对数据集进行数据清洗；
2：设置固定格式的提示模板；
3：读取数据集的每条数据，添加提示模板格式；
4：对每种情绪类别分别构建映射词表；
5：对数据集进行划分等处理；
6：模型超参数设置；
7：加载 BERT 预训练模型；
8：模型训练和验证；
9：保存模型参数；
10：在测试集上评估模型；
11：算法终止，流程结束。

2.3 实验与分析

2.3.1 实验设置

本项目的情绪分类任务采用的数据集来自两个数据集，分别是 NLPCC 2013 会议的情绪分类数据集和 OCEMOTION 中文情绪分类数据集。NLPCC 2013 情感分类数据集是 NLPCC 2013 官方测评任务的开源数据集，而 OCEMOTION 中文情绪分类数据集是来自于 NLP 预训练模型的泛化能力挑战赛的公开数据集。

本节实验中将两个数据集进行格式统一、数据整合筛选，得到新的情绪分类数据集，共有 54714 条数据，各情绪类别数据情况如表 3.3 所示，各情绪语句分布如图 2.3 所示。

表 2.3 情绪分类数据集各类别样本数量

情绪类别	数量

悲伤	9136
愤怒	8136
高兴	9788
喜好	8082
厌恶	8694
惊奇	5798
恐惧	5080

接着进行情绪分类数据集的划分。首先将数据集的数据随机进行打乱，接着按照 8：1：1 的比例对情绪分类数据集进行划分，分别为训练集、测试集、验证集，由此得到训练集样本数为 43772，测试集样本数为 5471，验证集样本数为 5471。

本次实验的实验环境为个人 PC 和服务器两种设备。个人 PC 配置了 AMD R7 5800H CPU 和 NVIDIA RTX3050 GPU，操作系统为 Windows 10 系统；服务器配置了 Intel i7 11800H CPU 和 NVIDIA GTX2080Ti GPU，操作系统为 Ubuntu 系统。另外，两者均配置了 Python 编程环境和深度学习框架。

本部分的实验包括模型训练等都在个人 PC 和服务器上进行，并且选择 PyTorch 框架来进行模型的建立、配置和训练等工作。

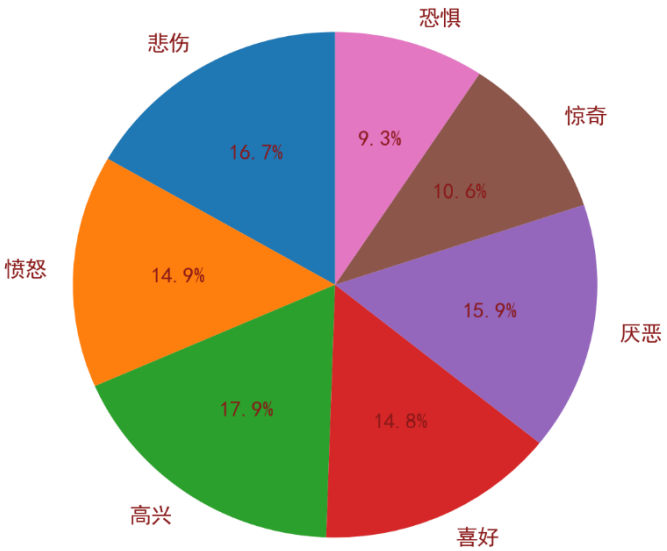

图 2.3 情绪分类数据集各类别样本分布

2.3.2 评价指标

本项目针对情绪分类器模型进行的实验，可以看作对于本项目情绪分类方法的准确性进行评估，分类越准确则可以证明情绪分类器的应用效果越好，则可得出模型性能越好。本项目选用了 *Accuracy*、*Recall*、*F1* 这三种常用的模型评价指标对分类效果进行评价。几种评价指标的含义如下：

（1）*Accuracy*

Accuracy 即准确率，是指用预测正确的样本数除以样本总数。*Accuracy* 指标的计算公式见公式(2.1)。

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.1}$$

在公式之中，*TP* 是指预测样本为真，实际样本也为真的情况；*TN* 是指预测样本为假，实际样本也为假的情况；*FP* 是指预测样本为真，实际样本为假的情况；*FN* 是指预测样本为假，实际样本为真的情况。当数据集的各类数据相对平衡时，*Accuracy* 的效果较好。

（2）*Recall*

Recall 即召回率，指的是在总的样本中有多少实际上为真的样本被选中，它也被称作查全率。*Recall* 指标的计算公式见公式(2.2)。

$$Recall = \frac{TP}{TP+FN} \tag{2.2}$$

Recall 表示了在整个检测的最终结果中为真的部分占数据集中实际为真的比重，*Recall* 越高表征着模型可以预测到更多正确的结果。

（3）*F1*

F1 是指 *F-Measure* 中的特定参数 α 为 1 时的情况。*F1* 指标的计算公式见公式(2.3)。

$$F1 = \frac{2*Percision*Recall}{Percision+Recall} \tag{2.3}$$

在公式之中，*Percision* 是指精确度，含义是在预测为真的样本中，有多少是确实为真的。*F1* 指标综合地考虑了精确度和召回率这两个因素，起到了对两者进行调和的作用。

2.3.3 实验结果与分析

本项目设计并进行了对比实验，在构建的情绪分类数据集上进行多次实验，以本研究的模型和其他几个基线模型进行对比，并通过多个评价指标进行定量评估验证，实验结果如表 2.4 所示。

表 2.4 实验结果

模型	*Accuracy*	*Recall*	*F1*
TextCNN	0.495	0.486	0.489
Bi-LSTM	0.623	0.615	0.598
BERT Fine-Tuning	0.672	0.685	0.680
BERT Prompt-Tuning	**0.719**	**0.732**	**0.716**

由表中的实验数据可得，Bi-LSTM 模型优于 TextCNN，下游微调的 BERT 模型优于 Bi-LSTM，经过提示微调的 BERT 模型优于正常下游微调的 BERT 模型，证明了 BERT 模型这种基于 Transformer 的语言模型拥有更加良好的特征抽取的能力，同时，提示学习能够更好地激发 BERT 预训练模型的潜能，使其在预训练阶段学习到的知识得到更充分的利用，让模型在情绪分类任务上表现更好。

基于以上结论，本项目选用在几个模型中情绪分类准确率最高的基于提示微调的 BERT 模型作为本项目的用户话语情绪分析模型，对用户输入的话语进行情绪分类，分类结果输入回复话语情绪预测模型中；同时本模型也运用在共情回复生成模型的情绪选择步骤中。

3 回复话语情绪预测

3.1 问题描述

本项目提出基于情绪知识图谱的回复话语情绪预测，因此通过对情绪知识图谱的构建，从而实现回复话语情绪预测功能。本项目的问题可以描述为由用户话语情绪分析模型分析得到的用户情绪为 e，通过情绪知识图谱对 e 所对应实体进行共情关系的查询预测，得到回复生成话语应该使用的情绪 r，且回复情绪 r 共有七种，分别是悲伤、高兴、难过、愤怒、喜欢、惊奇和恐惧。

3.2 情绪知识图谱设计

3.2.1 情绪领域本体构建

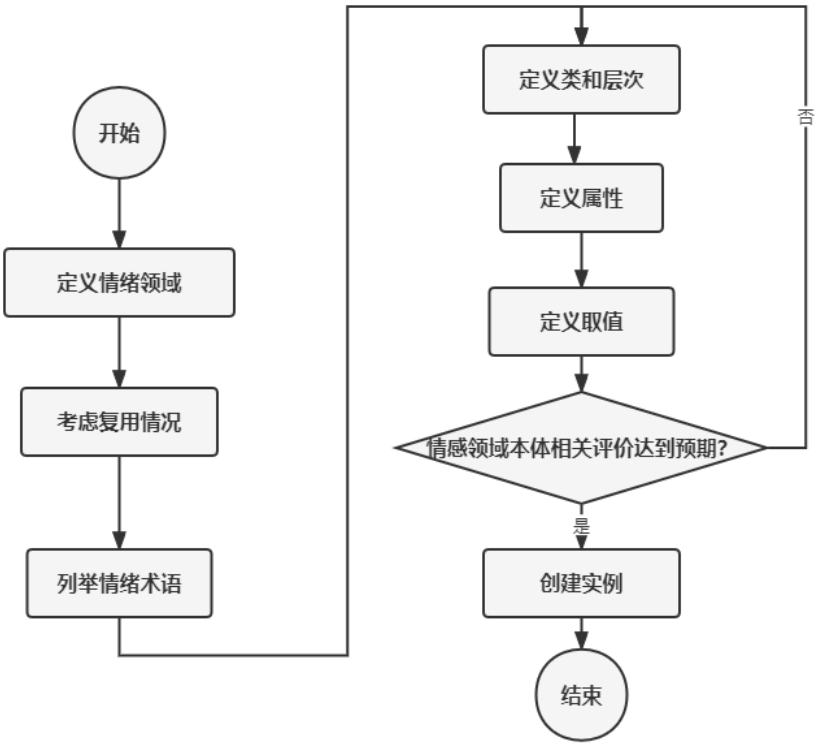

图 3.1 七步法构建情绪领域本体流程图

本体模型使通过概念的特征来抽象出的模型，它主要由类、概念、关系、实例等部分构成，通常可以使用四元组的形式进行表示。本项目构建的情绪领域本体可以由四元组 *EMO=(H,J,K,L)*来进行表示，其中 *EMO* 用来表示情绪领域本体；*H* 用来表示模型中领域的集合，由情绪类别、情绪词汇的各层次构成；*J* 用来表示模型中的关系，例如情绪之间相互的关系等；*K* 用来表示模型约束的集合，例如基本情绪产生几种派生情绪等；*L* 用来表示模型中的实体集合，例如情绪类别悲伤、愤怒等。

由此可知，本项目的本体模型涉及到了一些心理学的知识，同时涉及情绪领域的知识，实现情绪互相之间和情绪与情绪词汇之间的相互对应。

对于本体构建这一工程，现有的方法有 Methontology 方法、TOVE 法、七步法等，本项目选取七步法来进行本体的构建，七步法过程较为简单明了，适合

本项目的情绪领域本体。七步法构建情绪领域本体的流程如图 3.1 所示。

3.2.2 情绪知识图谱层次设计

基于前文设计的情绪领域本体，本项目进行了知识图谱层次的设计，具体层次如图 3.2 所示。

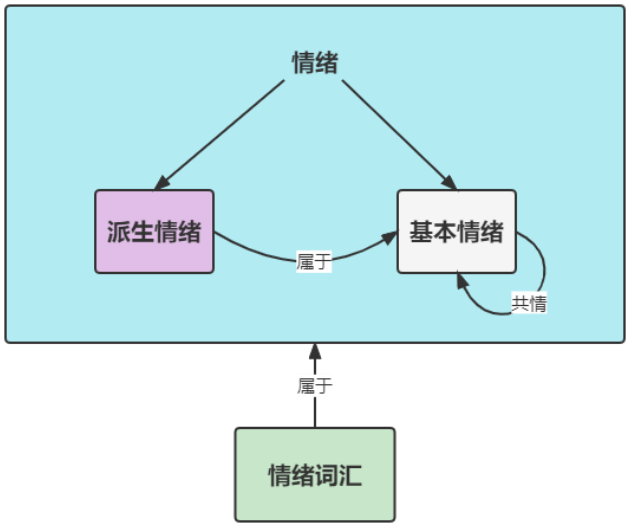

图 3.2 情绪知识图谱层次示意图

本项目充分利用了知识图谱中图网络的结构特点，对情绪类别之间的关系进行定义，为回复情绪预测功能提供了支持。本项目的情绪类别共分为两类，一类是基本情绪，另一类是派生情绪。派生情绪是由基本情绪细化而得来的，基本情绪共有七种，而派生情绪共有二十一种，每个基本情绪与一个或多个派生情绪相对应。

同时，本项目采用大量情绪词汇作为对话序列与情绪概念上的对应，使抽象的情绪更加具象化，并根据情绪词汇的同义词进行扩充，扩大知识图谱的规模。

3.2.3 情绪知识图谱属性定义

本项目通过对实体的属性进行定义来进行不同实体的区别，同一类别中的不同实体可通过其属性值的不同进行区别。

本项目中的两种情绪类别的属性表示相同，具体描述如表 3.1 所示。

表 3.1 情绪属性及其描述

属性	描述
名称	情绪名称
情绪类型	基本情绪/派生情绪
倾向	积极（1）、中立（0）、消极（-1）

本项目的情绪词汇的属性具体描述如表 3.2 所示。

表 3.2 情绪词汇属性及其描述

属性	描述
名称	情绪词汇名称
词性	情绪词汇的词性
强度	情绪词汇的情绪强度

3.3 情绪知识抽取和融合

本项目的情绪知识抽取包括知识获取和知识抽取。目前，情绪有关的知识目前来源少，获取难度较大，因此本项目采用了由大连理工大学研究出的中文情感词汇本体库，情感词汇本体库以 Ekman 模型为基础，将情绪分为七个种类，并对情绪大类进行了细化，同时收集了大量与情绪类别相对应的情绪词汇进行词性、极性、强度的人工标注，对本项目的情绪知识图谱构建有着不小的帮助。中文情感词汇本体库的一些示例如表 3.3 所示。

表 3.3 情绪词汇本体库示例

词汇	词性	词义号	词义数	极性	强度	情绪分类
批评	verb	1	1	5	2	NN
漂流	verb	2	1	5	0	PK

由于许多情绪词汇存在同义词，因此，本项目通过互联网上的百度同义词汇库选择了一些情绪词汇的同义词，将其极性、强度等与原词进行相同的设置，通过同义词库与原情绪词汇本体库的融合，得到扩充的情绪知识数据，用于情绪知识图谱的构建。

本项目的情绪知识抽取和融合的过程如图 3.3 所示。

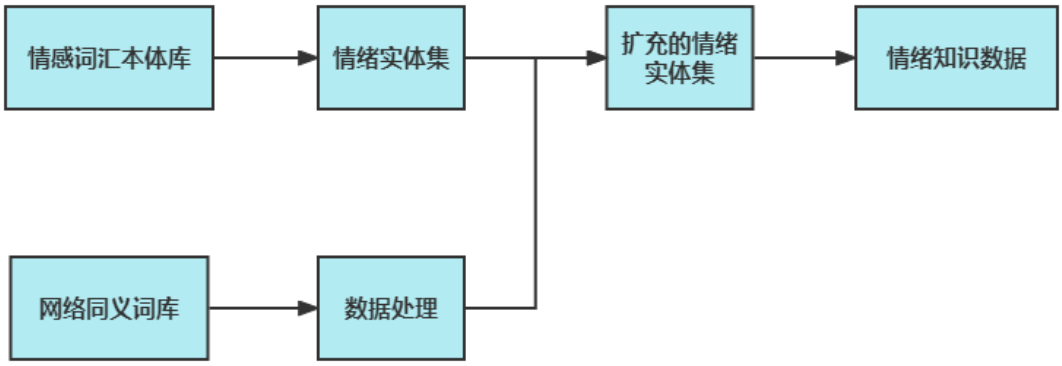

图 3.3 情绪知识抽取融合过程示意图

3.4 情绪知识图谱存储与可视化

本项目对构建的知识图谱进行存储，选用了 Neo4j 图数据库。Neo4j 数据库是一个高性能的非关系型图数据库，它不像传统的数据库那样把数据存储在表中，而是把数据存储在网络中。Neo4j 数据库可以被看作一个高性能的基于图的引擎，且具备一般成熟数据库的基本特性。Neo4j 数据库基于 JAVA 语言进行开发，它最初的设计目的是为了实现对于实体之间的描述达到更加高效的效果，传统的关系型数据库着眼于描绘实体本身的属性，它们互相之间的关系以外键来实现，因此对于关系的求解较为耗时。着重描述实体间关系的图数据库由此而生，以 Neo4j 为代表的图数据库在目前互联网呈爆发式增长的时代具有重要的应用价值。

本项目对于 Neo4j 图数据库的各种操作通过 Cypher 语言来进行，Cypher 语言作为一种图形查询语言，其使用方便且功能强，降低了使用难度。Cypher 语言的着眼点是在图中的找回，而不是操作，这样的设计使用户着重考虑查询的优化，淡化了实现细节。

本项目规定了存入 Neo4j 数据库的格式，实体数据格式如表 3.4 所示，关系数据如表 3.5 所示。

表 3.4 实体数据格式

字段	实体	属性	值
例 1	心虚	词性	形容词
例 2	失望	类型	派生情绪

表 3.5 实体间关系数据格式

字段	实体 1	关系	实体 2
例 1	喜好	共情	高兴
例 2	悲伤	属于	失望

首先使用 CSV 格式的文件对情绪实体与关系进行存储，并通过 Cypher 语言的批量导入并创建的语法将所有数据导入 Neo4j 数据库，完成情绪知识图谱的构建。情绪知识图谱的局部图如图 3.4 所示。

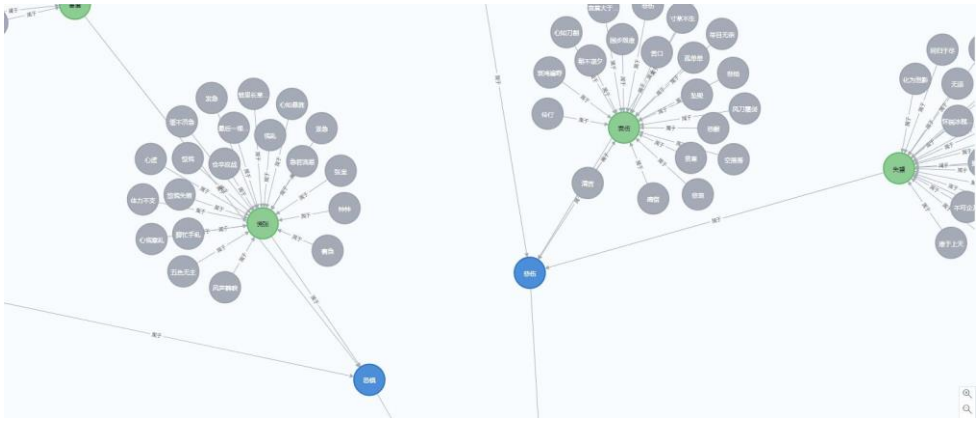

图 3.4 情绪知识图谱局部图

在对回复情绪进行预测时，首先通过用户话语情绪分析模型得到用户的情绪，然后使用 Cypher 语言的关系查询语法中的 MATCH 语法，对用户情绪所对应的情绪实体进行特定关系指向的情绪实体查询，从而得到预测的回复话语情绪，实现回复情绪预测的功能。

回复话语情绪预测的算法流程如表 3.6 所示。

表 3.6 回复话语情绪预测算法流程

输入：用户话语的情绪类别编号；
输出：回复话语情绪；
1：读取用户话语的情绪类别编号；
2：通过映射关系得到用户的基本情绪名称；
3：连接 Neo4j 图数据库；
4：执行 Cypher 查询语句；
5：返回查询结果；
6：将查询结果保存，作为回复生成模型输入的一部分；
7：算法终止，流程结束。

3.5 实验与分析

（1）情绪知识图谱规模

本项目的情绪知识图谱总体上包括了各类情绪和它们对应的情绪词汇，根据基本情绪进行分类的各类别实体数量如图 3.5 所示。总体上，本情绪知识图谱的实体总数为 42324。

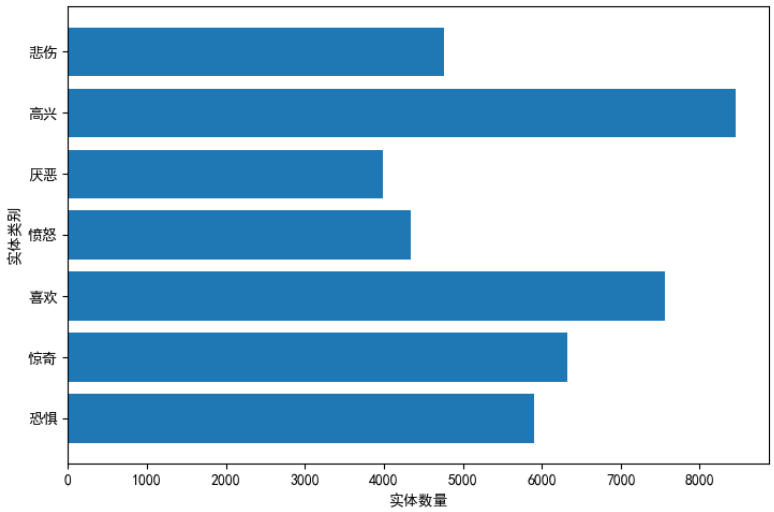

图 3.5 情绪知识图谱的基本情绪实体分布

本项目的情绪知识图谱中的关系包括基本情绪-基本情绪、基本情绪-派生情绪、派生情绪-情绪词汇三类。基本情绪-基本情绪关系的数量为 7，基本情绪-派生情绪关系的数量为 21，派生情绪-情绪词汇关系的数量为 42296。

（2）情绪知识图谱测试

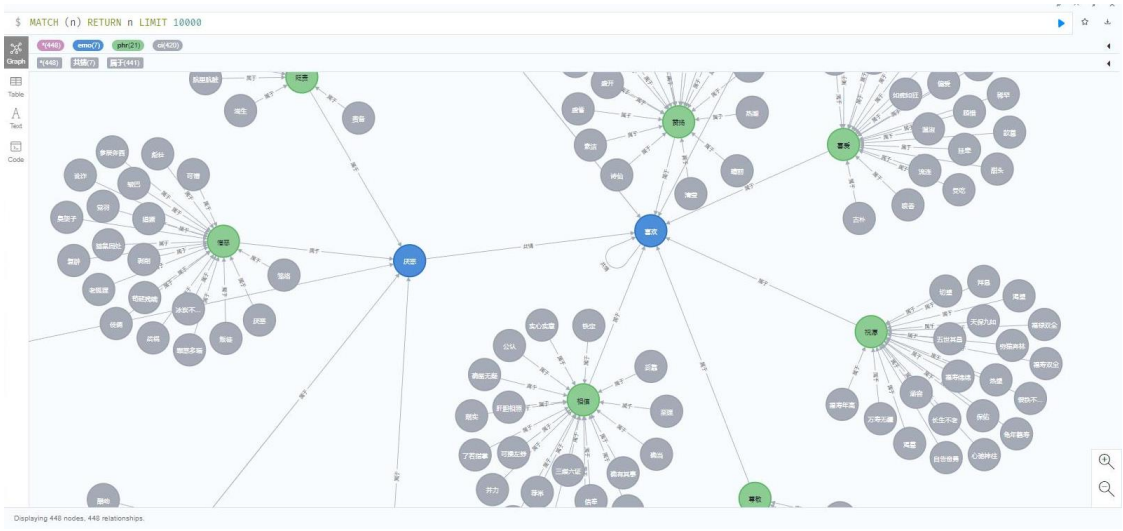

图 3.6 情绪知识图谱可视化

首先通过个人 PC 运行命令提示符，进入 Neo4j 存储的 bin 目录后，输入 neo4j.bat console 命令运行 Neo4j，输入 localhost 链接后进入 Neo4j 界面，可以看到情绪知识图谱可视化示意图如图 3.6 所示。

运行通过 Python 语言编写的用知识图谱查询进行回复情绪预测的文件，通过调用开源库 py2neo，用 Cypher 语言中的 MATCH 语法进行预测操作，根据情绪知识图谱的基本情绪之间的共情关系由情绪“惊奇”预测得到回复情绪“高兴”的运行结果如图 3.7 所示。

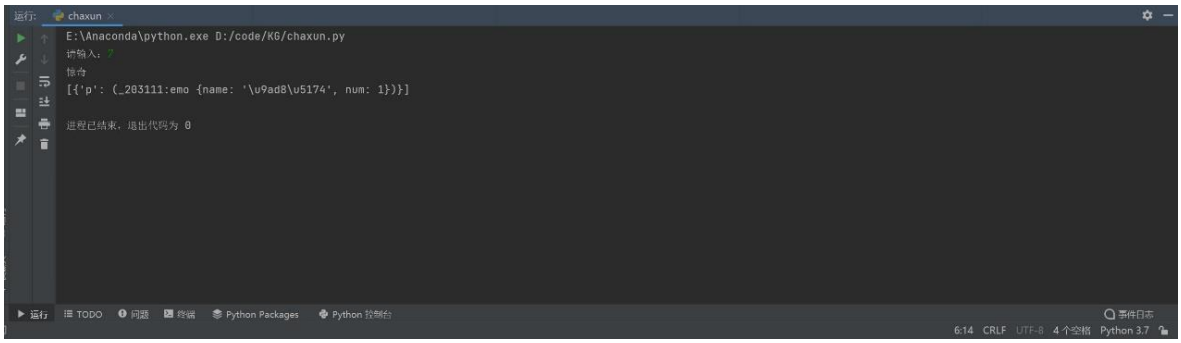

图 3.7 情绪知识图谱预测结果图

4 共情回复话语生成

4.1 问题描述

本项目主要的研究目标是共情对话生成任务，基于前面两章节用户话语情绪分析和回复情绪预测而得出的回复情绪，使对话生成模型能够生产含有该类情绪的对话，并通过情绪原因抽取来优化共情机制，从而生成共情对话。

共情回复话语生成任务可以用以下文本来描述：用户输入话语的文本词序列为$A = (a_1, a_2, \dots, a_x, \dots, a_n)$，通过共情知识图谱预测得到的回复情绪为 r，对话生成模型需要根据此输入的对话来生成相对应的回复语句 $B = (b_1, b_2, \dots, b_y, \dots, b_m)$。在这其中，$a_x$的意义是输入话语序列的第 x 个单词，输入话语序列的长度为 n，b_y的意义是回复话语序列的第 y 个单词，回复话语序列的长度为 m，回复情绪 r 共有七种，分别是悲伤、高兴、难过、愤怒、喜欢、惊奇和恐惧。

4.2 共情回复生成模型

4.2.1 总体架构

共情回复生成任务的最终目的是使模型能够生成指定类型情绪的回复，并达到与用户的共情，因此本项目设计的共情回复生成模型（ERGM, Empathic Reaction Generation Model）主要包括情绪原因抽取单元、GPT-2 回复生成模型、MMI 模型、情绪选择单元四个部分。情绪原因抽取单元主要用于得到情绪原因句并与原句进行拼接；GPT-2 回复生成模型主要用于候选回复的生成；MMI 模型主要用于提升模型生成的多样性；情绪选择单元主要用于使模型最终输出与预测情绪一致的回复。

共情回复生成模型的详细工作流程如下：

（1）将用户当前输入的语句、用户历史输入、预测完成的回复情绪作为模型整体的输入；

（2）情绪原因抽取单元将用户当前对话与历史输入作为子句建模为一个文档，并抽取当前对话的原因句与当前对话进行拼接；

（3）将拼接得到的句子输入 GPT-2 模型，模型最终输出 8 个候选回复句；

（4）MMI 模型将求得每个候选回复的 loss，按从小到大对候选回复进行排序；

（5）情绪选择单元调用第三章训练好的情绪分类器模型对每个候选回复进行排序，并选择与预测得到的回复情绪一致的回复中排序最靠前的回复作为最终的共情回复。

总的来说，本项目的共情回复生成模型总体架构如图 4.1 所示。

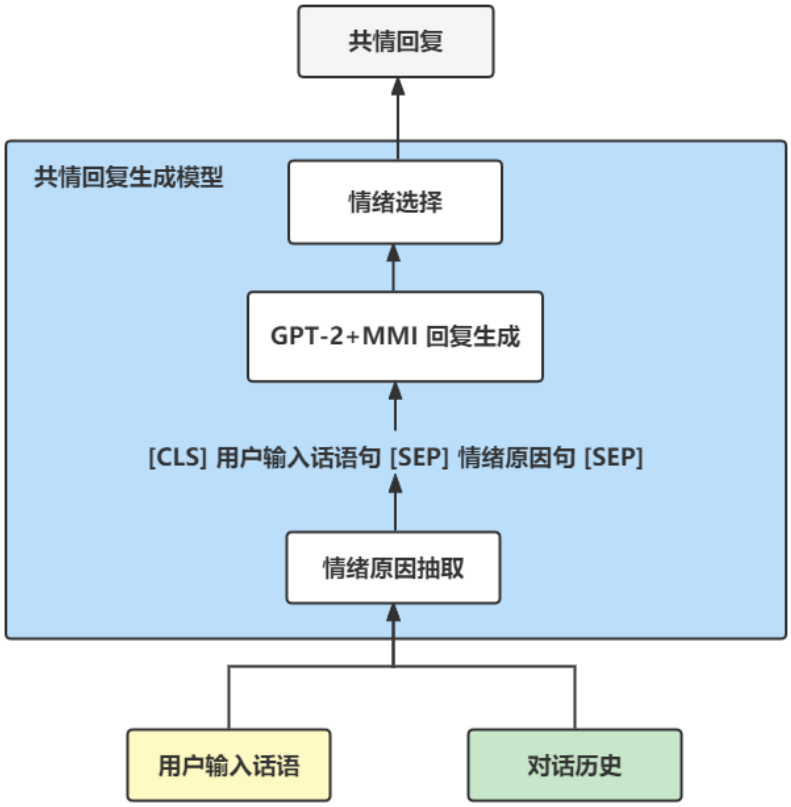

图 4.1 共情回复生成模型总体架构图

4.2.2 情绪原因抽取单元

情绪原因抽取是指将用户本次输入话语与用户历史输入的九句话一起建模为一个具有多个子句的文档，通过抽取文档中的情绪-原因对（ECPE）来判断用户当前话语是否具有情绪原因，若有的话则将情绪原因句与当前话语进行拼接，输入回复生成模型中。

本项目选取了 E2EECPE 模型来实现情绪原因抽取的功能。首先，将用户本次输入话语与用户历史输入的九句话一起进行合并，分别作为子句 c_i建模为文档 D，D 的表示见公式(4.1)。

$$D = \{c_1, c_2, \dots, c_i, \dots, c_m\} \tag{4.1}$$

其中，c_i 是指建模形成地文档 D 的内部子句，m 是指总的子句数，且满足 $1 \leq m \leq 10$,子句 c_i可以进一步分解为单词，其表示见公式(4.2)。

$$c_i = \{w_1, w_2, \dots, w_i, \dots, w_n\} \tag{4.2}$$

其中，n 是指句子中的单词数。总的来说，E2EECPE 模型最终会从 D 中提

取一组情绪-原因对 P，其表示见公式(4.3)。

$$P = \{(c_k^e, c_k^c)\} \tag{4.3}$$

在此公式中，c_k^e 和 c_k^c 分别代表最终得出的第 k 对中的情绪句和原因句。E2EECPE 模型架构如图 4.2 所示。

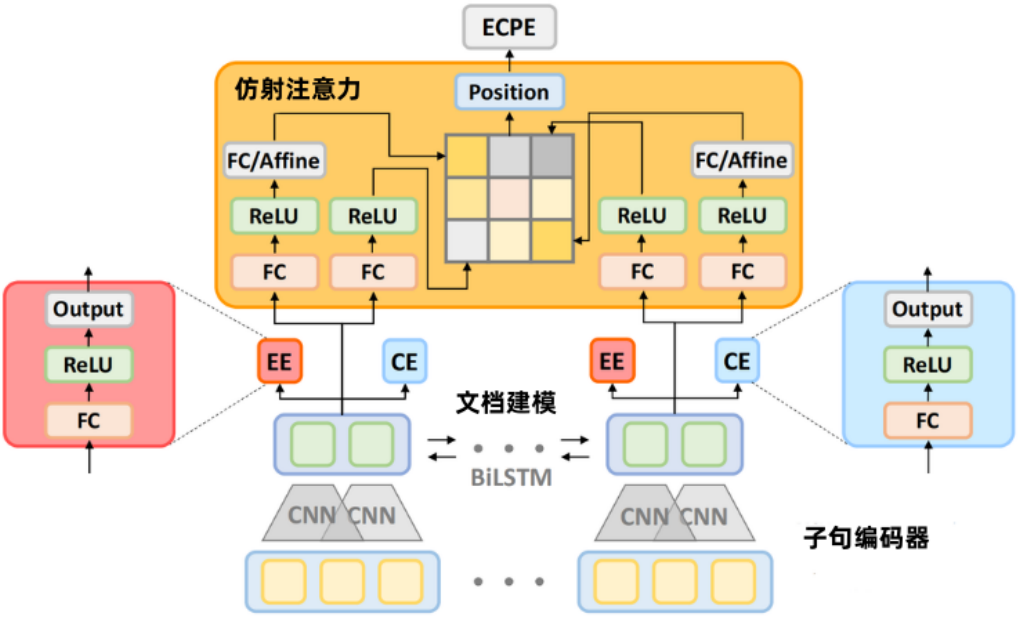

图 4.2 E2EECPE 模型架构图

E2EECPE 模型首先通过使用 TextCNN 模型给子句进行编码，将每个单词使用 Word2vec 表示为低维向量，然后通过使用 TextCNN 模型来获得子句的向量表示，进而使用最大池化进行特征的提取，最终得到具有上下文关系的子句向量。

通过 TextCNN 得到的子句向量不含其他子句的信息，E2EECPE 模型接着通过 Bi-LSTM 对子句进行建模，进行全文档 D 的建模。

最后，E2EECPE 模型使用了双仿射注意力机制，将文档 D 看作一张图，子句作为图的节点，直接对节点之间的相似度得分进行计算，判断两个子句之间是否构成情绪原因对。

通过调用 E2EECPE 模型对此时用户输入话语以及历史话语建模成的文档进行情绪-原因对提取之后，判断提取的情绪原因对是否有此时用户输入作为情绪句的情绪-原因对，若有的话则将对应的原因句与本句进行拼接，作为 GPT-2 对话生成模型的输入，此过程将情绪原因进行融入，增加了共情程度。

4.2.3 对话生成

本项目的对话生成通过使用预训练模型在构建好的对话数据集上训练的方式进行实现。在预训练模型的选择上，本项目选取了 OpenAI 公司使用大规模语料进行无监督训练得到的开源语言模型 GPT-2 作为预训练模型，其模型的实现是基于 HuggingFace 公司的 Transformers。

GPT-2 模型沿用了 GPT 单向 Transformer 的特点，在预训练的阶段增加了语料的规模，使训练语料的领域更广泛，提升模型的泛化能力，同时增加了 Transformer 的堆叠层数，使模型参数量增大。GPT-2 摒弃了模型 fine-tuning 的操作，不必对不同类型的任务进行下游微调，而是通过模型自身的自动识别进行任务的判断，对于本项目的对话生成任务较为适合。

本项目在 GPT-2 模型的基础上，通过自主构建的开源对话语料对 GPT-2 模型进行自回归训练，将数据集每次对话中的所有文本序列进行拼接，合成一个较长的文本，以固定的结束符作为结尾，将 GPT-2 预训练模型的任务定义为生成任务。GPT-2 预训练模型用于对话生成任务的模型结构如图 4.3 所示。

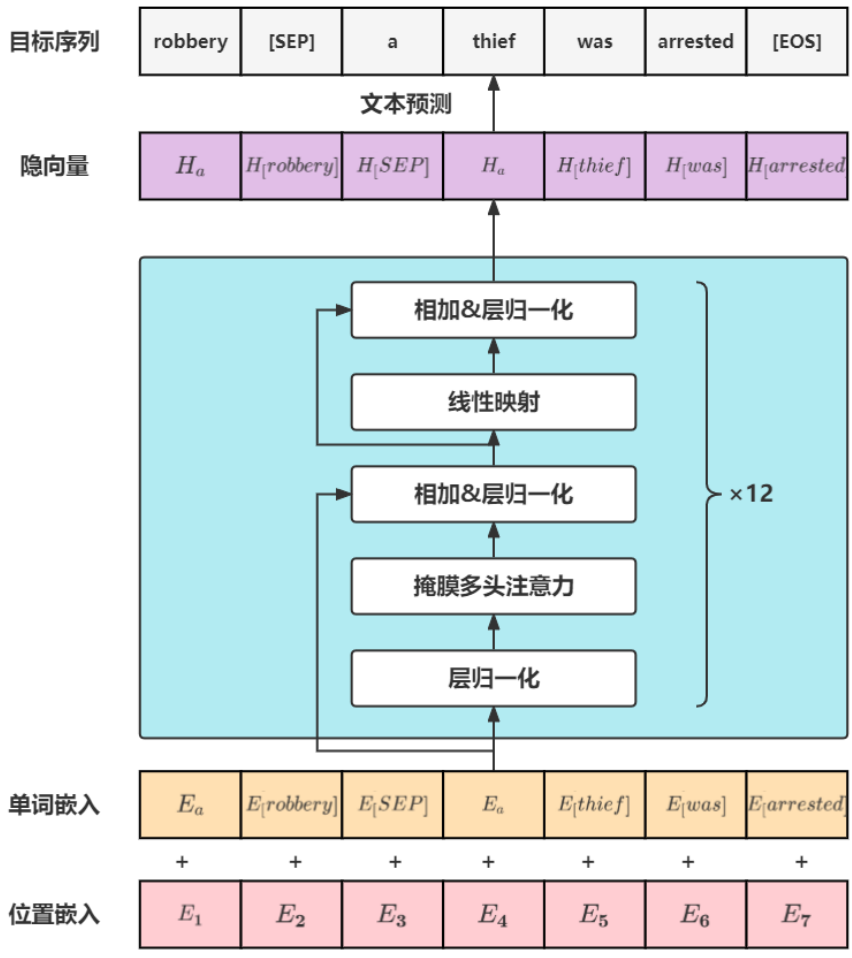

图 4.3 GPT-2 模型进行对话生成示意图

本项目将数据集中每次进行的多轮对话转换为一个长文本 s，此长文本序列的表示可见公式(4.4)。

$$s = m_1, m_2, \dots, m_x \tag{4.4}$$

其中，$m_1, m_2, \dots, m_x$为长文本序列的单词，x 为长文本序列的长度。因此，m_x和对话回复生成目标 t 的表示可见公式(4.5)。

$$t = m_{x+1}, m_{x+2}, \dots, m_y \tag{4.5}$$

$P(t|s)$的条件概率的表示可见公式(4.6)。

$$P(t|s) = \prod_{k=x+1}^{y} P(m_k|m_1, m_2, \dots, m_{k-1}) \tag{4.6}$$

由此可见，$P(t|s)$的条件概率能够表示为一系列条件概率的乘积形式。

另外，在 GPT-2 预训练模型用于文本生成任务时，解码过程中采用了集束搜索（Beam search）算法的解码模式，这种方法常常会出现输出文本不连贯或循环往复的效果。

这种现象是由于最大似然使解码进行循环反馈导致的，使用最大似然进行解码在分布上与人类文本差异巨大，因此会使解码结果出现不连贯或循环的情况，这种情况对于对话回复生成来说有着不利的影响。因此，本项目采用 Nucleus Sampling 采样方法来对 GPT-2 模型进行回复生成解码的优化。

Nucleus Sampling 即 Top-p Sampling，它是对 Top-k Sampling 采样方法的一种改进，Nucleus Sampling 不再是选定一个固定的 k 值，而是给定一个概率的阈值 p，定义一个最小集合为 V_p，且 V_p 在用于解码的候选词语 x 的集合当中，同时使候选词最小集合的所有概率之和大于或等于 p，其表示可见公式(4.7)。

$$\sum_{x \in V_p} P(x|x_{1:i-1}) \geq p \tag{4.7}$$

Nucleus Sampling 的好处是随着用于解码的词语集合的概率分布变化以及时间步的变化，候选词语的集合会发生动态的变化，而不是一个固定的区间。动态变化调整能够使解码生成的语句更加通顺，并且多样性得到了一定的优化。

最终改进的 GPT-2 模型通过解码生成 p 个候选回复句，此时生成的句子未进行情绪类型的约束，需要将候选句送入 MMI 模型中进行排序，然后再通过情绪分类器进行情绪选择。

4.2.4 最大互信息优化

本项目通过引入最大互信息（MMI，Maximum Mutual Information），对 GPT-2 模型的对话生成进行优化。GPT-2 模型进行对话生产任务的下游微调后，其输出文本有着经常为安全回复的现象，安全回复是指“嗯嗯”、“好的”这类概率较高的回复。本项目在共情回复生成模型中加入了互信息，采用了最大互信息的策略，降低模型输出安全回复的概率。

通过对话数据集进行训练的 GPT-2 模型原来仅会生成一个回复，本项目的上一小节对其进行改进，使 GPT-2 模型在输出时进行 Nucleus 抽样，生成 p 个回复作为候选回复，同时引入了 MMI 模型，使用候选回复与前文句子的关联概率来对候选回复进行排序。加入最大互信息模型之后回复生成效果如图 4.4 所示。

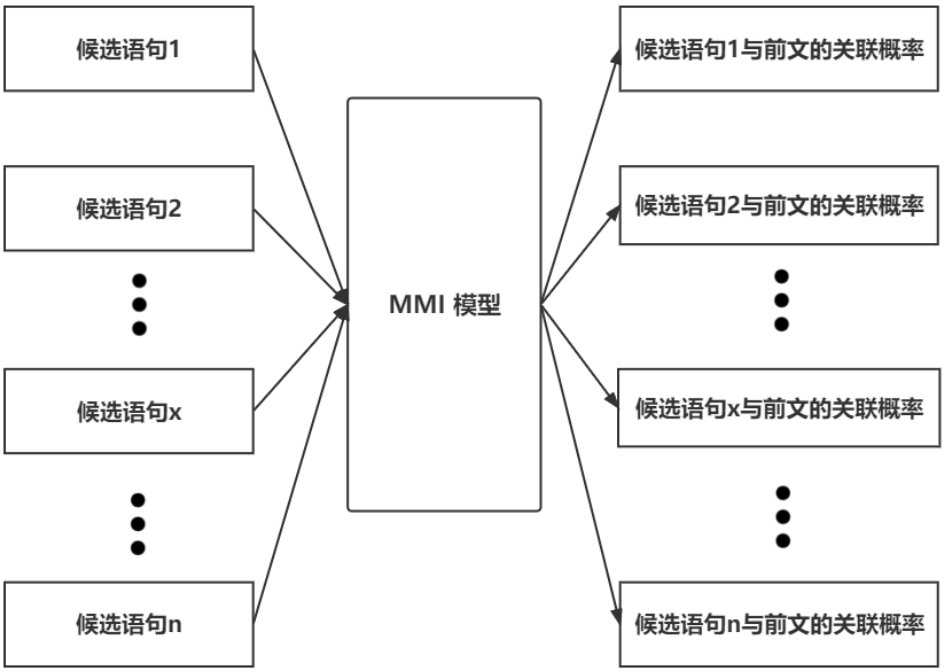

图 4.4 最大互信息模型效果图

最大互信息模型将 GPT-2 模型生成的每个候选对话与前面的对话历史按从后向前的顺序进行拼接，通过拼接的序列来计算该候选对话与前文句子的关联概率，并按概率从高到低进行排列，然后进行下一步的情绪选择操作。

因此，接近安全回复的候选语句会被最大互信息模型进行惩罚，因为经常重复的候选语句会出现与很多可能存在的查询产生关联的情况，这样以来降低了特定情况下出现查询的概率。

由此可见，最大互信息模型的评价函数所使用的评分机制，较好地解决了对话生成模型经常生成安全回复的问题，使本项目的回复生成效果更好。

4.2.5 情绪选择单元

本项目采用情绪选择策略，通过对 GPT-2 模型生成的若干句回复调用情绪分类器模型进行情绪分类，并根据预测得到的回复情绪进行情绪选择，然后根据每个候选回复的互信息排序决定最终的输出回复。

情绪选择策略的具体流程如下：

（1）读取 GPT-2 模型生成并经过 MMI 模型排序的 p 个候选对话，调用前文的情绪分类器得到每句话的情绪类别标签。

（2）读取通过情绪知识图谱预测得到的回复情绪。

（3）根据预测的回复情绪对生成的 p 个候选对话进行筛选，选择符合的若干个候选对话。

（4）选择符合的若干个候选对话中排序最高的作为最终的回复话语，完成情绪选择过程。

情绪选择单元的使用使模型最终输出的语句蕴含了预测得到的回复情绪，经过情绪选择步骤得到了模型的最终输出的共情回复。

总的来说，共情回复话语生成模型的算法流程如表 4.1 所示。

表 4.1 共情回复话语生成算法流程

输入：用户输入话语，用户历史输入，预测得到的回复情绪；
输出：共情回复话语；
1：读取用户输入话语和历史输入；
2：输入话语和历史输入建模为多个子句的文档 D
3：情绪原因抽取单元从文档 D 中抽取情绪原因句；
4：情绪原因句与目前输入句按格式进行拼接；
5：GPT-2 模型读取拼接后的语句序列；
6：生成 8 个候选回复句；
7：MMI 读取候选回复句，求得每句话与前文句子的关联概率；
8：将概率按从高到低进行排列；
9：情绪选择单元读取候选回复句排序的结果；
10：调用情绪分类器对候选回复句进行分类；
11：从分类结果中保留符合预测回复情绪的候选回复句；

12：选择目前候选回复句中关联概率最高的作为共情回复句；

13：输出共情回复话语；

14：算法终止，流程结束。

4.3 实验与分析

4.3.1 实验设置

本项目的共情对话生成任务采用的数据集选取了自然语言处理领域中常用的开源对话数据集，包括青云语料库和豆瓣多轮对话数据集两种。首先将两份数据集进行合成，并通过数据清洗操作对一些特殊符号以及其他噪声进行预处理，最终得到约 50 万多轮对话语料作为数据集，接着按照 8：1：1 的比例对数据集进行划分，分别为训练集、测试集、验证集。

本部分的实验环境和实验设备与前文相同，因此对这部分不再赘述。

4.3.2 评价指标

本项目针对共情对话生成模型进行的实验，可以看作对于本项目共情对话生成方法的生成效果而进行评估，生成的话语共情效果和多样性越好则可以证明共情对话生成模型的应用效果越好，则可得出模型性能越好。本项目节通过使用两种方法对共情回复生成模型进行评价，分别是自动评估和人工评估，两种方法的评价指标如下：

（1）自动评估指标

本项目对于共情对话模型的自动评估的指标包括 *PPL* 和 *Distinct-1/2*。

1）*PPL*

PPL 是指困惑度（Perplexity），它计算的基本思想是模型训练完后，其在测试集上的概率值越高，意味着模型效果越好。

假设句子 *s* 是由单词构成的序列，其表示形式见公式(4.8)。

$$s = w_1, w_2, \dots, w_k \tag{4.8}$$

句子 *s* 的概率计算公式见公式(4.9)。

$$P(s) = P(w_1, w_2, \dots, w_k) = P(w_1)P(w_2|w_1) \dots P(w_k|w_1, w_2, \dots, w_{k-1}) \tag{4.9}$$

因此，*PPL* 的计算公式可见公式(4.10)。

$$PPL(s) = P(w_1, w_2, \dots, w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_2, \dots, w_N)}} \tag{4.10}$$

在公式中，计算句子 *s* 的概率，并取其倒数后开根号，说明句子的概率越大，*PPL* 困惑度的值越小，则模型效果越好。

2）*BLEU*

BLEU 是指双语评估替补（Bilingual Evaluation Understudy），它可以用来计算模型生成语句与参考语句之间的相似程度，以此来评估一组自然语言任务生成文本的效果，*BLEU* 的计算公式见公式(4.11)。

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} W_n \times logP_n\right) \tag{4.11}$$

在此公式中，*BP* 指的是惩罚因子，*W* 表示正权重，*N* 表示 *n-gram* 的总长度，*P* 表示改进的 *n-gram* 精度值，*n* 为当前匹配 *n-gram* 的长度。*BP* 对预测句长度小于真实句长度的情况进行惩罚，其计算公式见公式(4.12)。

$$BP = \begin{cases} 1 & if\ lc > lr \\ e^{(1-\frac{lr}{lc})} & if\ lc \leq lr \end{cases} \tag{4.12}$$

在此公式中，lc 为候选生成语料库的总长度，lr 表示语料库的有效参考长度。

3）*Distinct-1/2*

本项目的共情回复模型属于对话生成模型，需要对生成结果进行多样性的判断衡量。*Distinct* 指标是一种专门用来评估序列多样性的指标，它的定义见公式(6.13)。

$$Distinct(n) = \frac{Count(unique\ ngram)}{Count(word)} \tag{4.13}$$

在此公式中，*Count(unique ngram)*代表了序列中非重复的 *ngram* 的数量，而 *Count(word)* 代表了序列中 *ngram* 的总计数值。*Distinct-n* 数值越大表征了序列的多样性程度越高。

（2）人工评估指标

通过自动评估能够显示模型的性能，但是不能替代人的感受来评判回复的情绪表达等效果，因此本项目设置了人工评估。本项目的人工评估指标结合了本项目的研究目标和通用的对话生成质量评测的任务标准，设置了以下几个指标：

1）情绪引导

此指标表征了对话能否引发对话用户的情感共鸣。评价标准：0，完全不能；1，稍微可以；2，很可能。

2）情绪表达

此指标表征了对话能否表达出态度和适当的情绪。评价标准：0，无；1，稍有；2，很明显。

3）逻辑关联

此指标表征了对话与上文的主题是否有关联性。评价标准：0，完全无；2，稍有；3，非常有。

4）对话持续

此指标表征了对话能否对与用户对话的持续起到促进作用。评价标准：0，完全不能；1，稍微可以；2，很可能。

4.3.3 实验结果与分析

（1）自动评估结果

本项目的自动评估实验在对话数据集上进行，将本项目的共情对话生成模型（ERGM）与目前对话生成任务主流的模型 Seq2Seq 模型等进行对比，同时通过删除情绪原因抽取单元和最大互信息优化模块进行了消融实验研究，并通过多个自动评价指标进行评估验证，实验结果如表 4.2、表 4.3 所示。

由表中的实验数据可得，本项目提出的共情对话生成模型（ERGM）在四项自动评价指标中相对于其他模型均取得了最好的得分，证明了 ERGM 在对话生成效果方面具有较好的性能，同时证明了情绪原因和最大互信息对于模型的重要性。

表 4.2 自动评估结果

模型	*PPL*	*BLEU*	*Distinct-1*	*Distinct-2*
Seq2Seq	183.0	1.34	0.0082	0.0304
Seq2Seq+Att	169.3	1.77	0.0101	0.0342
GPT-2	141.9	2.11	0.0127	0.0401
CDial-GPT	149.3	2.39	0.0104	0.0396
ERGM	**137.6**	**2.42**	**0.0134**	**0.0425**

表 4.3 消融实验结果

模型	*PPL*	*BLEU*	*Distinct-1*	*Distinct-2*
ERGM	**137.6**	**2.42**	**0.0134**	**0.0425**
w/o E2EECPE	141.2	2.01	0.0122	0.0366
w/o MMI	144.5	2.15	0.0115	0.0389

（2）人工评估结果

本项目在对模型的人工评估方面，选取了 50 组对话结果作为样本，并随机选取了三位评测人员根据上文中的评分规则进行打分，并对打分结果采用归一化方法进行处理，实验结果如表 4.3 所示。

表 4.4 人工评估结果

模型	情绪引导	情绪表达	逻辑关联	对话持续
Seq2Seq	0.185	0.431	0.290	0.315
Seq2Seq+Att	0.365	0.445	0.410	0.486
GPT-2	0.405	0.524	0.542	**0.630**
CDial-GPT	0.402	0.566	0.512	0.577
ERGM	**0.467**	**0.612**	**0.592**	0.609

由表中的实验数据可得，本项目提出的共情对话生成模型（ERGM）在四项人工评价指标的前三项中取得了最好的得分，在第四项指标中取得了较好的得分，证明了 ERGM 在共情方面具有较好的效果。

（3）结果示例

本项目对共情对话生成模型进行对话效果测试，得到了一些典型的具有共情效果的测试对话，结果的两个示例如表 4.5 和表 5.6 所示。

表 4.5 结果示例 1

用户话语输入	共情话语输出
今天的电影好搞笑	哈哈，我也是这么想的
你也看了吗	是呀是呀
今天好高兴	哈哈，为什么呀
因为今天中了彩票	哈哈你在逗我吗

表 4.6 结果示例 2

用户话语输入	共情话语输出
你好呀	你好你好
今天的天好阴沉呀	对呀，要下雨了
昨天考试挂科了	我也挂了呀
我哭了	不哭不哭，下次加油
这会好难受	抱抱你

5 共情对话原型系统的设计与实现

5.1 系统总体设计

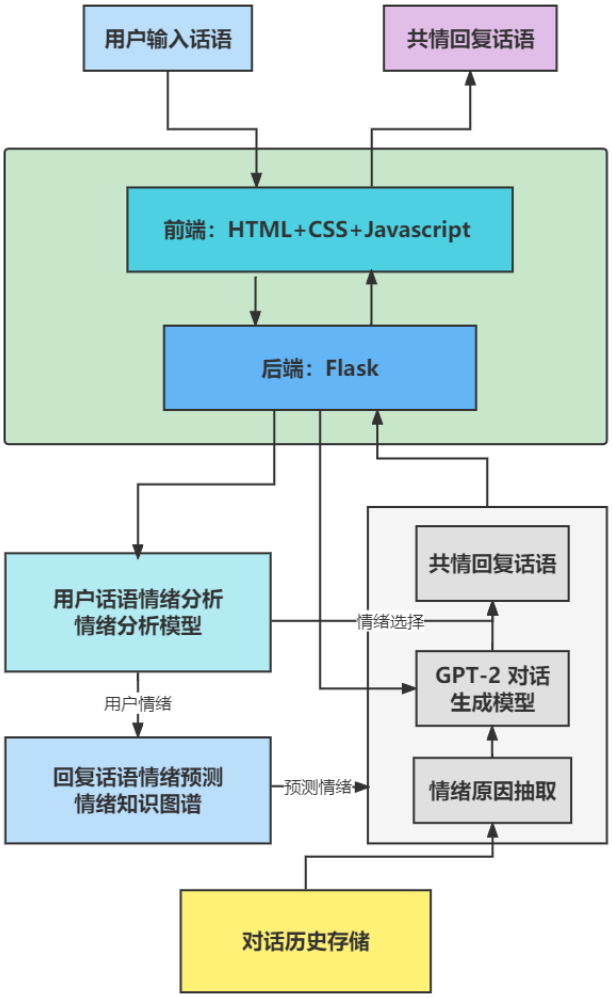

图 5.1 共情对话原型系统整体设计图

本项目主要进行共情对话原型系统的总体构建和测试。共情对话原型系统的主要功能是实现人机交互，结合了本项目的第二、第三、第四部分训练的模型和方法，同时考虑到系统的可重用性和可拓展性，设计开发出一套原型系统。

共情对话原型系统的呈现形式是一个文本输入、输出形式的人机对话 Web 页面，实现用户与聊天机器人的自然流畅的交互，完成了一个面向人机共情的对话生成原型系统。共情对话原型系统的整体设计方案如图 5.1 所示。

5.2 系统功能设计

本系统的功能主要分为以下几部分：

（1）人机交互功能

因为共情对话原型系统主要是主要面向用户进行使用，因此本项目考虑了系统与用户的交互，针对体验性和交互性进行了设计。

对话系统界面整体上是一种类似于社交对话窗口的形式，主要使用 Web 技术进行开发, 采用了 HTML、CSS、JavaScript 作为开发语言，HTML 主要作为对话网页内容的载体，CSS 的作用是作为对话页面的样式表现，JavaScript 的作用是实现对话网页的一些对话框等特殊效果；另外考虑到人性化，进行了 UI 美化，最终共情对话系统的交互界面可以运行在网页浏览器上。

（2）后台控制功能

本项目采用 Flask 框架进行系统后端的开发，进行系统的后台控制。Flask 是一种轻量级的 Web 开发框架，采用 Python 语言进行编写。Flask 框架具有很强的定制性能，且代码简洁易于实现，很适合作为本项目共情对话原型系统的开发框架。

（3）用户话语情绪分析功能

用户话语情绪分析功能主要通过第三章介绍的情绪分类器来实现。对于情绪分类模型的训练方面，首先整合了 7 分类的细粒度情绪分析数据集 OCEMOTION 和 NLPCC 2013 中文微博细粒度情绪分类数据集作为模型的数据集，并对数据集进行预处理，按照 8：1：1 的比例划分，作为模型的训练集、测试集和验证集；对 BERT 预训练模型进行下游微调，并采用提示学习构建模板，在合适的位置添加[MASK]符号，同时构建了情绪映射词表。最终训练得到的模型可实现通过情绪分类器分析用户输入话语中的情绪，对情绪进行分类。

（4）回复话语情绪预测功能

回复话语情绪预测功能主要通过第四章介绍的情绪知识图谱以及上一步分析出的用户情绪来综合实现。首先基于情绪心理学知识以及情感词汇本体库，构建面向共情关系的情绪知识图谱，并选择 Neo4j 数据库来进行知识图谱的存储与可视化，然后基于系统的用户情绪分析结果，通过情绪知识图谱进行 Cypher 查询来预测相应的回复情绪。

（5）共情回复话语生成功能

共情回复话语生成功能主要通过第五章介绍的共情回复生成模型并结合上一步预测得到的回复情绪以及情绪分类器来综合实现。对于共情回复生成模型的训练方面，首先选取了开源的聊天语料：青云语料库和豆瓣多轮对话数据集，并对语料进行数据预处理，作为本部分模型的训练集和测试集；选取 E2EECPE 模型进行情绪原因句的抽取，并将情绪原因句与当前输入句进行拼接；选取预训练模型 GPT-2 进行下游训练，将多轮对话语料建模为长文本，成为文本生成任务，对于用户输入话语可生成 8 个候选回复；同时在模型中加入互信息，使回复更具有多样性，避免出现安全回复。最后调用情绪分类器对候选回复进行情绪分类，从符合预测回复情绪的候选回复中选取互信息模型排序最高的作为对话回复，生成了最终的共情回复。

5.3 系统工作流程

本项目所设计开发的共情对话原型系统的目的是满足用户的情感和交流需求，因此共情对话原型系统需要具有较高的回复效率、对话质量和人机共情程度。

在对话质量方面本项目通过 Nucleus 采样和 MMI 模型提升了原型系统的对话质量；在共情方面，本项目通过分析用户话语情绪，通过知识图谱对相应的回复情绪进行预测，同时从对话历史中抽取情绪原因句，作为对话生成模型的一部分，最后通过情绪分类器对生成模型生成的候选回复进行情绪选择，这一系列过程满足了共情的需求。

共情对话原型系统的详细工作过程如下：

（1）等待用户进行话语的输入；

（2）首先得到用户输入话语，通过用户话语情绪分析模块的情绪分类器模型得到用户话语的情绪分类，将用户话语情绪输入回复情绪预测的情绪知识图谱

中，预测得到应该采用的回复情绪；

（3）本步骤与（2）同时进行。首先得到用户输入话语，通过情绪原因抽取单元在历史对话中抽取得到情绪原因句，将情绪原因句与用户输入话语句进行拼接输入 GPT-2 对话生成模型，得到 8 个候选回复句，最后将候选回复通过 MMI 模型进行排序；

（4）对 GPT-2 模型生成并根据互信息进行排序的每个候选回复调用情绪分类器确定其情绪类型，并根据预测得到的回复情绪进行候选回复的情绪选择，在符合的候选回复中选择互信息排序最高的候选回复作为最终的共情回复；

（5）通过前端页面输出共情回复；

（6）本轮对话完成，进行接下来的对话轮，等待用户再进行话语输入。

5.4 系统测试

5.4.1 系统测试环境

共情对话原型系统整体采用 Python 语言进行代码的编写，前端部分采用 HTML、CSS、JavaScript 作为开发语言。对于共情对话原型系统的运行和测试采用了 Linux 远程服务器，共情对话原型系统的运行环境硬件配置如表 5.1 所示。

表 5.1 原型系统运行环境硬件配置表

硬件类型	型号
处理器	Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz
内存	16GB 3200Mhz
显卡	NVIDIA GeForce GTX 2080Ti
硬盘	512G SSD

5.4.2 系统功能测试

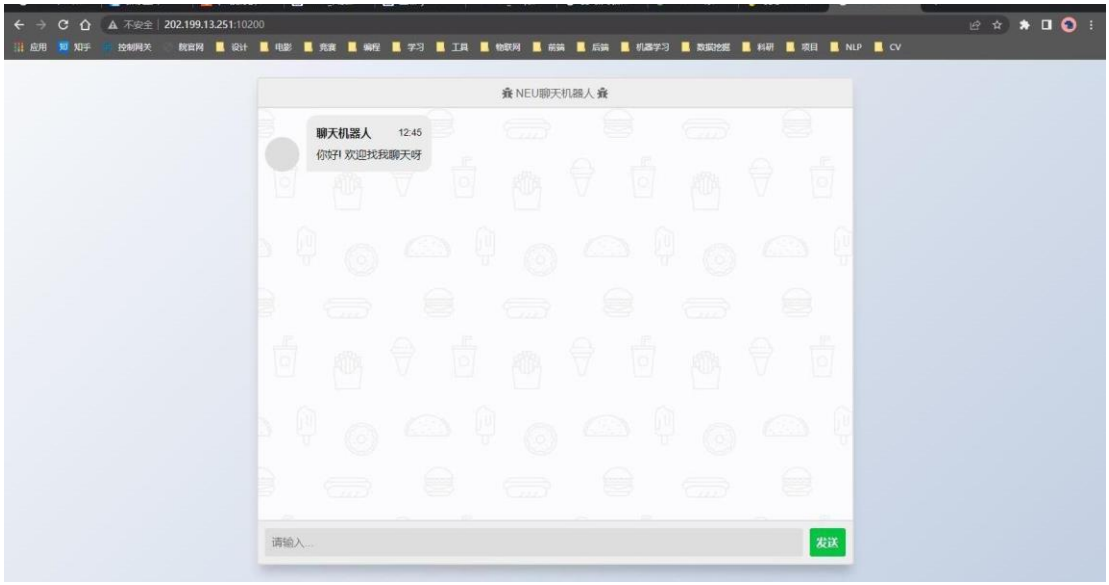

图 5.2 共情对话原型系统初始界面

首先通过 SSH 连接 Linux 远程服务器，打开 Flask 后端程序文件，运行该文件。通过本地主机浏览器打开网页链接，此时可以看到，本项目共情对话原型系统的初始界面如图 5.2 所示。

通过网页的聊天框输入文本，并点击发送键即可实现对话功能。本项目原型系统的聊天效果如图 5.3 所示。

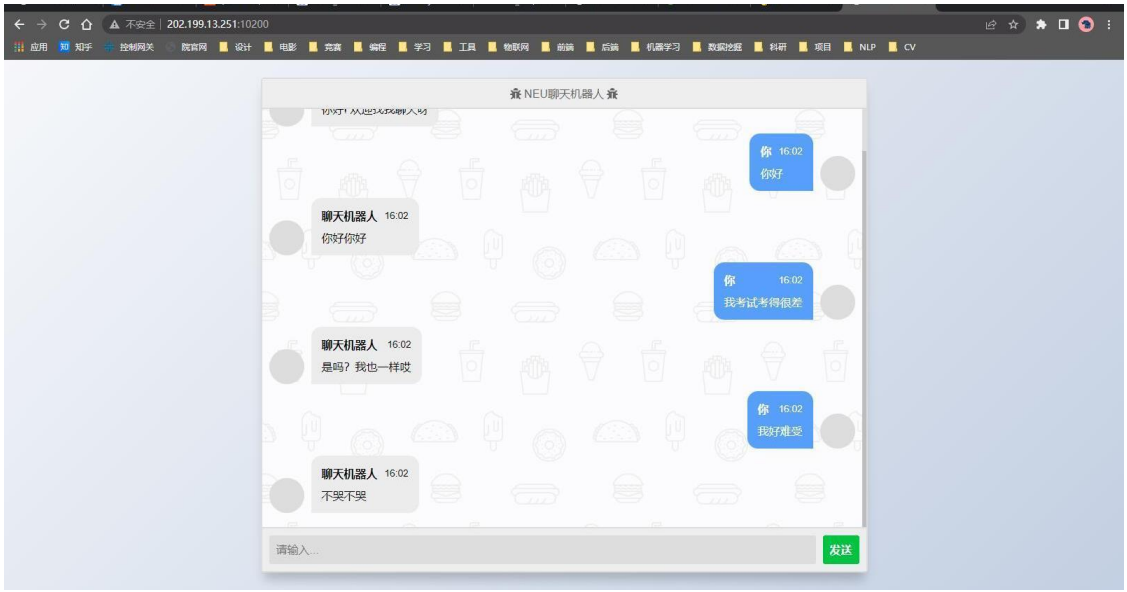

图 5.3 共情对话原型系统聊天效果图

6 实践项目总结

本次自然语言处理课程实践项目在之前对话生成以及情绪分析等领域研究成果的基础上，对面向人机共情的对话生成系统进行了研究和实现。

首先，本项目对用户话语情绪分类进行研究，基于 BERT 预训练模型和提示学习技术，通过构建特定的模板和情绪映射词表，建立了对 BERT 模型进行提示微调的情绪分类模型，并通过对比实验验证了情绪分类模型的有效性；然后，本项目对回复情绪预测进行研究，并通过情绪领域本体构建、情绪知识图谱层次设计、情绪知识图谱属性设计、情绪知识抽取和融合构建了情绪知识图谱，并对其进行存储与可视化；接着，本项目实现了共情回复话语的生成，结合情绪分类模型和预测的回复情绪，通过情绪原因抽取单元、GPT-2 回复生成模型、MMI 模型和情绪选择单元来构建一个共情回复生成模型，并通过自动评估和人工评估实验验证了模型对于共情和回复质量提高的有效性；最后，本项目在前几部分的基础上，基于 Web 技术实现了共情对话原型系统的构建和测试。

本项目的研究对于人机共情对话生成方面进行了一定的模型上的改进，取得了一些效果的提升，但由于研究精力、时间以及经验有限，本次项目在未来将会从多个方面继续进行改进，主要体现在如下方面：

（1）本项目的对话生成模型主要针对单轮对话的生成，没有对多轮对话的场景进行考虑，因此可能会发生对话上下文不一致的情况，不符合人们平时的交流方式，在未来有待于拓展为多轮对话模型。

（2）本项目的共情方式主要采用了用户话语情绪分析、回复话语情绪预测、情绪原因抽取融合、情绪选择等部分来实现，在未来有待于考虑实现更加完善的共情策略，比如情绪引导、定向安慰等策略对模型的人机共情效果进行提升。

（3）本项目的用户话语情绪分类研究部分基于对 BERT 预训练模型进行提示微调来进行，而本项目的研究只构建了一种提示学习模板，不同的模板对于分类的效果有着重要的影响，因此可以尝试构建更多的模板进行对比实验，探究更优的模板构建方式。

（4）本项目的情绪知识图谱中基本情绪之间的共情关系较为简单，有待于将来进一步研究，丰富情绪知识图谱的内容，并使共情关系更加合理。