基于提示学习的小样本文旅客服问题分类方法

薛赟琨[1],王　亮[2],朱欣娟[1]
(1. 西安工程大学 计算机科学学院,陕西 西安 710600;
2. 秦始皇帝陵博物院,陕西 西安 710699)

摘　要:为解决传统分类方法需要大量的标签数据,且在样本量不足的情况下模型容易过拟合的问题,提出了一种基于提示学习的小样本文旅客服问题分类方法。首先,收集整理了文旅客服领域数据,对问题类别进行了层次分类,构建了文旅客服问题数据集,并对其进行数据类别标注;其次,基于提示学习的方法,设计了八类不同的提示模板,分别在 Bert、RoBerta 等预训练语言模型上进行对比实验,选择了效果较优的提示模板和预训练语言模型,并对文本分类模型进行领域微调;最后,设计了基于提示学习的层次分类算法,将算法和微调后的分类模型应用于某博物馆文旅客服问答系统。实验结果表明,在小样本数据条件下,文旅客服问题分类的准确度和 F1 值达到 92.03% 和 96.43%,相较于 BERT 文本分类基线模型,准确度和 F1 值分别提升了 5.77 百分点和 5.78 百分点。

关键词:提示学习;问题分类;小样本;层次结构;文旅客服问答
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2025)05-0188-09
doi:10.20165/j.cnki.ISSN1673-629X.2024.0399

Prompt Learning Based Classification for Few-shot Cultural and Tourism Customer Service Questions

XUE Yun-kun[1], WANG Liang[2], ZHU Xin-juan[1]
(1. School of Computer Science and Technology, Xi'an Polytechnic University, Xi'an 710600, China;
2. Emperor Qinshihuang's Mausoleum Site Museum, Xi'an 710699, China)

Abstract: To solve the problem of traditional classification methods requiring a large amount of labeled data and model overfitting in the case of insufficient sample size, a few-shot cultural and tourism customer service question classification method based on prompt learning is proposed. Firstly, data in the field of cultural and tourism customer service was collected and organized, and question categories were hierarchically classified. A dataset of cultural and tourism customer service questions was constructed, and the data categories were annotated. Secondly, based on the method of prompt learning, eight different prompt templates were designed and compared on pre trained language models such as Bert and RoBerta. The better performing prompt templates and pretrained language models were selected, and domain fine-tuning was performed on the text classification model. Finally, a hierarchical classification algorithm based on prompt learning was designed, and the algorithm and the fine tuned classification model were applied to a museum cultural and tourism customer service Q&A system. The experimental results show that under few-shot data conditions, the accuracy and F1 score of cultural and tourism customer service question classification reach 92.03% and 96.43%. Compared with the BERT text classification baseline model, the accuracy and F1 score have increased by 5.77 percentage points and 5.78 percentage points.

Key words: prompt learning; question classification; few-shot; hierarchical structure; question answering with cultural and tourism customer service

0　引　言

随着游客数量的增长和旅游需求的多样化,传统的旅游服务模式已经难以满足游客的需求,而文旅客服问答系统的构建可以为游客提供更加便捷、高效和个性化的服务。问答系统按照答案获取的方法可以分为检索式问答和生成式问答[1]。在生成式问答中如 ChatGPT、ChatGLM、Llama 等大语言模型在文本生成、翻译、问答等任务上展现出卓越的性能,但会产生“幻

收稿日期:2024-10-13　　**修回日期**:2025-02-13
基金项目:陕西省重点研发计划项目(2024GX-YBXM-548);国家文物局 2023 年文物科学技术研究项目(2023ZCK026)
作者简介:薛赟琨(2000-),女,硕士研究生,研究方向为自然语言处理;通信作者:朱欣娟(1969-),女,教授,博士,CCF 专业会员(R4757M),研究方向为智能信息处理。

觉[2]”,即模型生成看似合理的内容,但内容是不正确的或与输入无关[3]。而且由于数据的质量和多样性存在偏差,且生成式模型是黑盒模型,导致泛化能力不足,在特定领域内存在可靠性和公正性的问题[4],例如 ChatGPT 在专业领域难以处理复杂的语言现象,导致在文本分类任务上效果仍低于微调模型[5]。在文旅领域,游客对景点信息、时间安排、票务查询等问题的答案要求高度的准确性,信息往往根据情况随时调整,生成式问答对信息的准确度存在局限性,因此检索式问答更适用于文旅客服场景。

检索式问答是在预先设定好的问答库中查找答案,可以提供准确的答案;主要由问题分析、信息检索和答案抽取组成[6],而问题分类是问题分析的重要任务。传统的问题分类模型构建需要大规模的标注数据,在样本量不足的情况下,训练模型容易过拟合,而特定领域内数据很难达到以上要求,人工标注成本昂贵且耗时[7];为解决以上问题,小样本学习[8]被提出,但是如何在小样本数量条件下,使得模型适应下游任务[9],仍是亟待解决的问题。基于生成式模型 GPT-3 的启发[10],将提示学习应用于小样本文本分类中,使得在样本数量少的情况下也能得到较好的分类结果。

文本分类旨在将未标注的句子分配到预定义的标签或类别中[11],随着自然语言处理(Natural Language Processing,NLP)技术的发展,其研究经历了四个阶段[12]。早期通常基于传统机器学习算法构建,常见的算法有贝叶斯、viterb 算法、SVM-KNN 算法[13]等,由于特征提取需要大量的人工标注工作,泛化能力较弱。随着 NLP 神经网络模型的出现,逐渐应用以神经网络为代表的模型架构,自动学习文本与任务之间的特征关系,但也需要大量的标注数据训练网络模型,常见的方法有 CNN、RNN、Seq2Seq 模型等。Lai 等[14]提出了循环卷积神经网络模型(Recurrent Convolutional Neural Networks,RCNN),它结合了 CNN 和 RNN,能很好地捕捉到文本序列中的上下文信息,提高分类性能,缺点是模型计算复杂度和训练资源需求较高。Chen[15]首次实现了用 CNN 进行句子分类的 TextCNN 模型,该模型特征提取能力强、计算效率高、模型结构简单,但对全局上下文信息和长距离词语之间的依赖关系处理不足,且在数据稀缺的情况下,模型泛化能力受限。

为了解决人工标注问题,预训练语言模型[16](Pretrained Language Model,PLM)被提出。PLM 基于大量无标注数据训练,针对特定的下游任务用少量的有标签数据微调训练后的 PLM,使其获得对应下游任务的语言表达能力[17]。Devlin 等[18]提出了预训练语言模型 Bert,它是基于 Transformer 架构的一种深度双向模型,通过在大规模文本语料上进行预训练,模型可以学习到丰富的语言表示,从而在 NLP 各类任务上取得优异的性能。随后 OpenAI 团队提出了生成式预训练语言模型 ChatGPT[19]系列,随着模型的迭代升级,训练数据规模也在不断的增强,模型参数规模也随之急剧增加,导致微调阶段需要大量的标注数据,才能保证不会发生训练过拟合和不稳定的问题。目前,“预训练、微调”的形式逐渐被“预训练、提示和预测”的形式所取代[20]。

提示学习能够充分激发预训练语言模型的潜能[21],所以被广泛应用于各个领域。Wallace 等[22]使用梯度搜索来激发 PLM 进行目标预测的短语。Schick 等[23]提出了提示学习的范式,介绍了一种新颖的半监督学习方法 PET,通过利用填空式问题来扩充训练数据,从而改善了在资源稀缺情况下的文本分类和自然语言推理任务的性能。Shin 等[24]提出了 Autoprompt 方法,可以自动生成提示且使用梯度引导搜索,为不同的领域任务构建提示,比手动构造模板更为高效。Zhong 等[25]提出了一种连续提示的 OPTIPROMPT 方法来优化提示。Li 等[26]提出了 prefix-tuning,冻结 PLM 的参数,但优化了与自然语言生成任务有关的特定向量,常用于生成任务。Liu 等[27]提出了 P-Tuning 模型,使用向量化的表示构建提示模板,将模板转化为参数优化的问题,实现了模板的自动构建,但 P-Tuning 模型只在第一层插入向量提示,其他层的提示都来自第一层,极大地约束了需要优化的参数量。为了改善上述问题,Liu 等[28]又提出了 P-Tuningv2 模型,在 P-Tuning 模型的基础上,对每一层句向量开头都插入向量的连续提示,改进了 P-Tuning 模型在难度高的任务上效果不好的问题。

经以上研究表明,提示学习的方法通过合适的提示将下游任务转化为适合模型的形式[29],充分挖掘模型的上下文处理能力,从而提升了 NLP 领域处理各种任务的效果。但是,目前提示学习应用在文旅领域的方法很少,并且没有提出贴合领域相关的提示模板。针对以上现状,该文对提示学习分类进行研究,提出了一种基于提示学习的小样本文旅客服问题分类方法(Prompt learning based Classification for Few-shot Cultural and Tourism customer service questions,PCFCT)。

论文主要贡献如下:

(1)创建了一个包含 3 000 条游客问题数据的文旅客服问答数据集,针对文旅客服问题的特点,设计了问题大类小类,并对数据集进行标注;

(2)提出了 PCFCT 方法,设计了一类适合文旅客服问题分类的提示模板,问题分类的准确度达到

92.54%;

(3)采用层次分类结构,将 PCFCT 方法应用于文旅客服问答系统,在小样本的场景下系统整体问答准确度达到 90.82%。

1 文中方法

1.1 数据集构建

该文基于某文旅机构提供的客服问答数据,通过人工标注的方法实现对问题数据的语料库标注,构建了文旅问答数据集。根据文旅客服数据的特点,对用户问题进行分类,设计了 32 个大类,其中 15 个大类下设小类,17 个大类直接对应问题的答案。图 1 展示了全部的大类和小类,虚线框中的加粗字体代表大类,虚线框中的实线框对应 15 个大类细化下的小类,由于大类问题占比不均匀,导致小类类别设计不均匀。

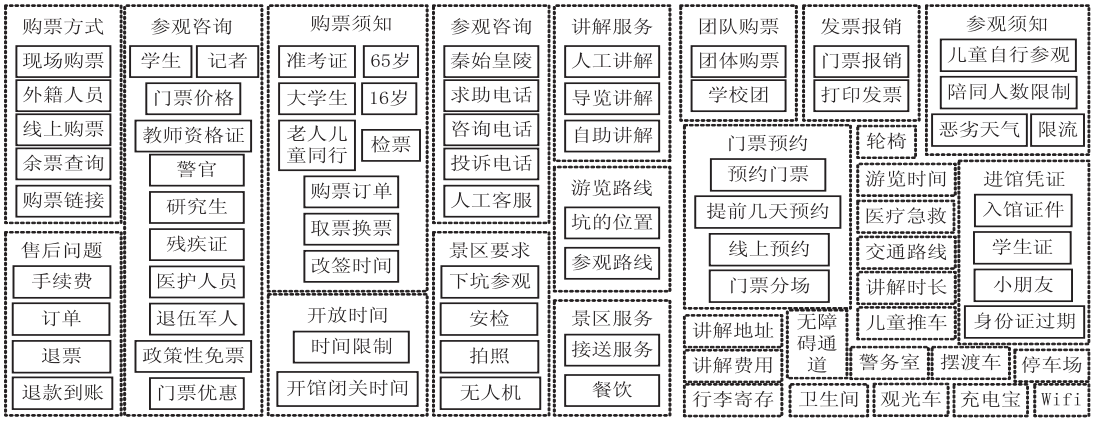

图 1 文旅客服问题大类和小类设计

具体标注流程:首先,基于设计的类别建立类别标签映射,即类别与数字标签的映射,人工对问题数据进行类别对应;其次,编写了数据处理的代码,主要对问题数据进行打标签和随机打乱数据的操作;最后,转换实验需要使用的文件格式。

该文采集了 1 960 条用户问题作为样本数据,因为问题数据类别占比不均匀,所以对数据进行适当扩充,最终用于问题大类分类的实验数据是 3 000 条;15 个需要细化小类的实验数据是从 3 000 条数据中选取出对应的问题,然后对小类问题数据进行处理和标注。该部分数据用于提示学习问题分类模块的训练和测试,实验均按照 8∶2 的比例随机划分训练集与测试集,图 2 展示了 32 个大类所占问题数量。

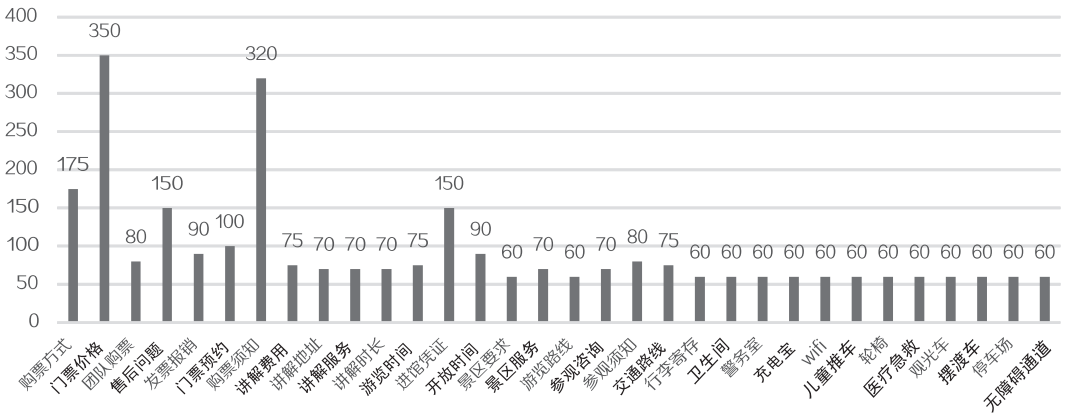

图 2 类别问题数量柱状图

基于上述问题数据,选取了包含类别关键词的问题,作为问答系统的标准问题,并根据基础数据以及某博物馆的相关文旅客服信息,设计了对应的标准答案,最终构建了标准问答数据库。表 1 展示了部分问题类别对应的标准问题及其标准答案,并将该部分数据用于最终问答系统的实现。

1.2 PCFCT 方法

在文本分类任务中,输入序列 $Q=\{q_0,q_1,\cdots,q_n\}$ 最终被划分到一个类标签 $z\in Z$ 中。而提示学习将分类任务转化为一个掩码预测的问题,PCFCT 方法结构如图 3 所示,提示模板是基于数据集特点进行设计,将提示模板加入到输入序列中,即原有的问题数据转变成加入提示模板的形式。这样可以使得 PLM 下游任务以掩码任务形式呈现。然后,通过 PLM 处理预测类别标签并填入[MASK]位置。最后,通过标签映射类别,即将预测到的标签映射到具体类别上。

表 1　部分问题类别及问答例句

大类	小类	标准问题	答案
购票方式	现场购票	可以现场买票吗	您好,进行票务购买,应统一在“兵马俑票务在线”公众号或小程序上预约买票
	外籍人员	外籍人员怎么购票	需要您自行到售票大厅去购票和咨询,祝您参馆愉快
	线上购票	可以线上(其它平台)购票吗	您可以在“兵马俑票务在线”公众号或小程序上预约购票,博物馆暂不与任何第三方售票平台合作
	余票查询	我马上就到景点了,今天还有余票吗	您好,可以打开官方微信公众号“秦始皇帝陵博物院”的购票界面,或“兵马俑票务在线”微信公众号及小程序,能够一目了然地看到每个时段的余票,购买相应时段的票并按时入园即可
	购票链接	怎样可以访问到购票链接	微信搜索公众号“兵马俑票务在线”,点击进入后,按照提示操作即可购票,预祝您参馆愉快
团队购票	团体购票	团队购票有优惠吗	很抱歉,博物馆没有团队票价
	学校团	有学校介绍信参加兵马俑研学的16 岁以下儿童团队该如何购票	该类问题您可以联系宣传教育部进行具体业务办理,电话 029-81399048 或 029-81399047
讲解地址		该如何寻求人工讲解服务	在博物馆大门售票厅的东侧,有讲解联系处,您可在此联系人工讲解
医疗急救		博物馆有医疗急救服务吗	有的,情况紧急,建议您立刻拨打电话,咨询电话:029-81399174
…	…	…	…

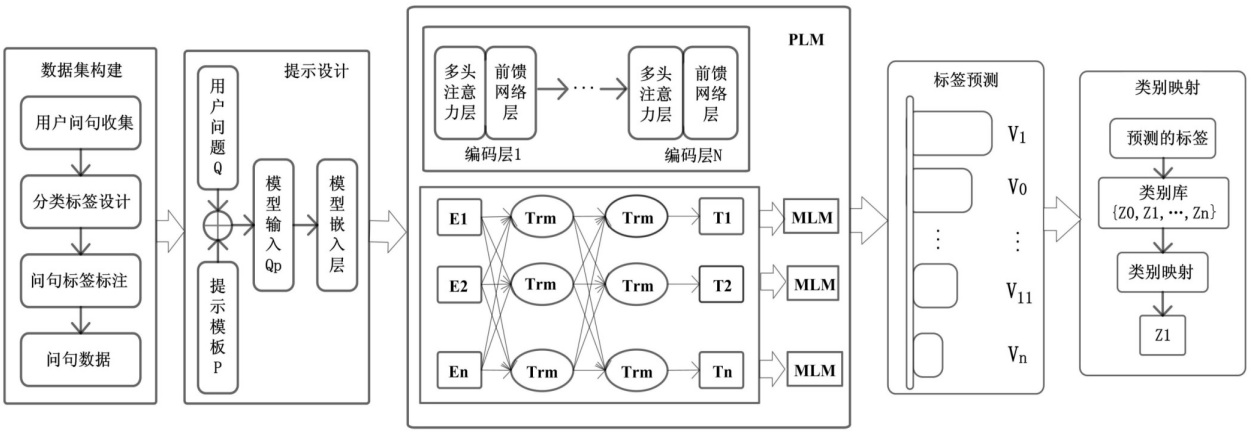

图 3　PCFCT 方法结构

1.2.1　构建提示模板与嵌入层交互

提示模板与模型嵌入层进行交互,即构建提示模板来影响输入序列的嵌入表示。提示学习最主要的任务就是构建提示模板,模板设计的好坏对实验结果有很大的影响。该文主要通过在问题输入时增加模板提示,将输入问题转化为加入提示的形式。假设需要对问题 Q =“退伍军人是否免票?”进行分类,将提示模板加入到输入序列中,模型输入序列即可表示为:Q_p =“退伍军人是否免票? 问题是[MASK]类。”在 PLM 的嵌入层中,提示模板会通过词嵌入映射到特定的向量空间,与原始问题输入序列的嵌入融合,形成上下文相关的向量表示。模板构建以及模型嵌入如图 4 所示。

该文设计了八类提示模板,<Query>是标准关键词,< Domain Keywords > 是领域关键词,< Task Keywords>是任务关键词,模板 1 和 5 包含标准关键词,模板 2 和 6 包含任务关键词,模板 3 和 7 包含领域关键词,模板 4 和 8 包含以上所有关键词;模板 1 ~ 4 是[MASK]位于句末,模板 5 ~ 8 是[MASK]位于句中。

表 2 中列出了在实验过程中使用的八类模板。P_1 ~ P_4 中[MASK]位于句末,P_5 ~ P_8 中[MASK]位于句中;P_1、P_5 属于包含标准关键词且提示文本长度较短的模板;P_2、P_6 包含任务关键词,例如类别、类等;P_3、P_7 包含领域关键词,例如文旅问答、景区等;P_4 是[MASK]位于句末且包含领域和任务关键词的提示模版;P_8 是[MASK]位于句中且包含领域和任务关键词的提示模版。P_1 ~ P_8 是大类使用的模板,用来验证模

板设计的合理性和可用性,同时对大类进行分类。

表 2 提示模板例句

序号	提示模板文本
P_1	问题属于[MASK]
P_2	用户问题属于类别[MASK]
P_3	这是文旅问答中的[MASK]
P_4	这类文旅客服问题是属于[MASK]
P_5	这是[MASK]问题
P_6	该句属于[MASK]类的问题
P_7	文旅客服问答中的[MASK]问题
P_8	文旅用户问题属于[MASK]类型的问题

1.2.2 标签预测

提示模板改变了 PLM 的注意力机制对输入序列的权重分布,提示中的向量会优先吸引注意力,指导 PLM 将更多的计算资源分配到与分类任务相关的部分。模型会根据提示模板提供的上下文信息生成概率分布,预测出最符合上下文的类别词。

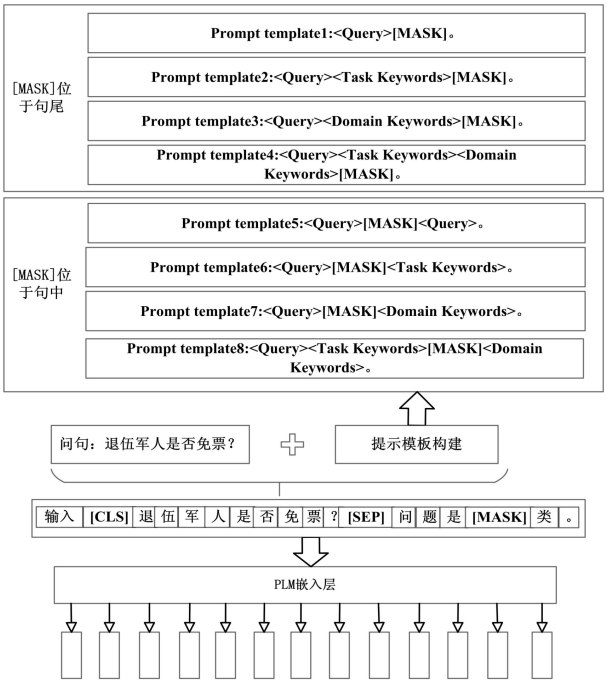

图 4 提示模板设计

PLM 主要是利用上下文信息预测出[MASK]位置概率最大的标签。通过 PLM 的嵌入层对输入序列进行编码,从而获得向量表示,然后进入模型编码器进行语义信息的提取,具体的 PLM 编码器由多个编码层组成,每个编码层都包括多头注意力层和前馈网络层,能够高效地捕捉到不同位置之间的依赖关系,最终利用掩码语言模型来理解和预测[MASK]位置的正确标签。具体标签预测过程如下:给定一组问题数据,输入序列记为 $Q=\{q_0,q_1,\cdots,q_n\}$,这些问题都对应到一个特定的类别 $z\in Z$ 中,设类别标签的集合为 $V_z=\{v_0,v_1,\cdots,v_n\}$。将待预测的问题加入提示后进行预处理,提取特征向量后输入到训练好的模型,模型计算每个标签被填入[MASK]中的概率为 $P([\text{MASK}]=v\in V_z\mid Q_p)$,然后选择概率最高的标签作为预测标签,将问题分类任务变成计算问题标签概率的任务,如公式 1 所示。

$$P(z\in Z\mid Q)=P([\text{MASK}]=v\in V_z\mid Q_p) \quad (1)$$

例如,对于上面问题模板 Q_p 的分类,如果 Q_p 属于 $z_1=v_1$“门票价格”类别的概率大于 $z_2=v_5$“门票预约”类别的概率,则该问题会被分类到标签为 v_1 的类别中。

1.2.3 类别映射

类别映射主要是将上一环节模型预测输出的标签映射到预定义类别上。具体实现是将预测到的标签概率计算好之后,将概率最高的标签作为预测结果映射到类别名称上。若要对预测结果与原始类别进行对应,该文采用标签得分加权平均函数来得到预测类别的分数,如公式 2 所示,得分最高的类别即为最终结果。

$$P(\hat{z})=\arg\max_{z\in Z}\left(\frac{\sum_{v\in V_z}P([\text{MASK}]=v\mid Q_p)}{|V_z|}\right) \quad (2)$$

其中,Z 表示问题类别,V_z 表示集合中的标签词,Q_p 表示问题加入提示模板后的输入序列。该文在计算预测问题标签 v 和真实问题类别 z 之间的差距时,运用交叉熵损失函数进行模型的训练,如公式 3 所示。

$$\text{loss}=-\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{c=1}^{C}\text{label}_i^c\log(\text{pred}_i^c)\right) \quad (3)$$

其中,N 是问题数量,C 是类别数量,label_i^c 是第 i 个问题的第 c 个类别的真实标签,pred_i^c 是模型对第 i 个问题的第 c 个类别的预测概率。

1.3 基于层次分类算法构建问答系统

1.3.1 层次分类算法

为了将分类任务的复杂性简化,系统采用层次体

系结构,通过 n 次分类逐渐缩小用户问题与标准问题的范围,可以有效防止数据不均匀对分类的影响。当用户提出领域相关的问题时,首先,调用第一层中类别的训练参数文件进行初次分类,分到该层具体类别之后,判断该类别下是否包含其他类别,若不包含,则直接把该类别对应的问题答案输出返回给用户;若包含其他类别,则调用该类别的训练参数文件进行二次分类,分类到具体类别之后,重复上述操作,直到该类别下未划分子类别。变量设置如下:共分为 K 层,每层包含多个类别,例如, L_1_category = $\{v_1, v_2, \cdots, v_n\}$, L_2_category = $\{v_1: \{v_{11}, v_{12}, \cdots, v_{1h}\}, v_2: \{v_{21}, v_{22}, \cdots, v_{2i}\}, \cdots, v_n: \{v_{n1}, v_{n2}, \cdots, v_{nj}\}\}$,以此类推设计 L_k _categor 类别;针对上述类别设计,层次分类算法的描述如下:

算法:层次分类算法

输入:用户问题输入序列 Query;

初始化:类别对应答案文件库 Answer(v);初始化总层数为 k ,加载 PCFCT 方法训练的类别模型参数函数的条件如下:

$$\text{Path}(v) = \begin{cases} L_1\_\text{path}(Q), K = 1 \\ L_k\_v_k\_\text{path}(Q), K = k \end{cases}$$

输出:分类后的类别结果 v 以及问题的标准答案 Answer(v)。

1:for Q in Query do:　//处理用户问句

2: $v_i = L_1$_path(Q) //对用户问题进行初次分类

3: print v_i　//输出当前类别

4: for k in K do:　//进入分类循环

5:　if v_i in L_k _category: //判断是否细化

6:　$v_{k_i} = L_k\_v_i$ _path(Q) //进入 L_k 层进行 k 次分类

7:　print v_{k_i}　//输出当前类别

8:　$k = k + 1$　//循环变量自增,转向步骤 4

9:　else:

10:　return Answer(v_{k_i})　//返回类别对应的答案

11:　end if

12: end for

13:end for

1.3.2　构建问答系统

基于 PCFCT 和层次分类算法构建某博物馆文旅客服问答原型系统,根据文旅客服领域的数据特点划分为两层,每层有多个类别;构建问答系统整体流程如图 5 所示。

对于相似性较高的问题,例如:“ Q_1:研究生买票有优惠吗?”“ Q_2:高中生买票有优惠吗?”“ Q_3:与 65 岁以上老人该如何买票?”, Q_1、Q_2 属于门票价格类, Q_3 属于购票方式类,但 Q_1 属于门票价格类中的研究生类, Q_2 属于门票价格类中的学生类, Q_3 属于购票方式类中的 65 岁类别,这样进行详细的层次划分,可以有效地避免相似问题混淆;文中问答系统做两次分类将任务分解为多个类别训练模块,每个模块可以独立训练和优化,防止问题分类混乱,提高最终问答的准确性。由于类别标准问答库中的标准问题都对应着唯一的标准答案,所以最终返回给用户的答案是划分到具体类别中最符合用户意图的答案。

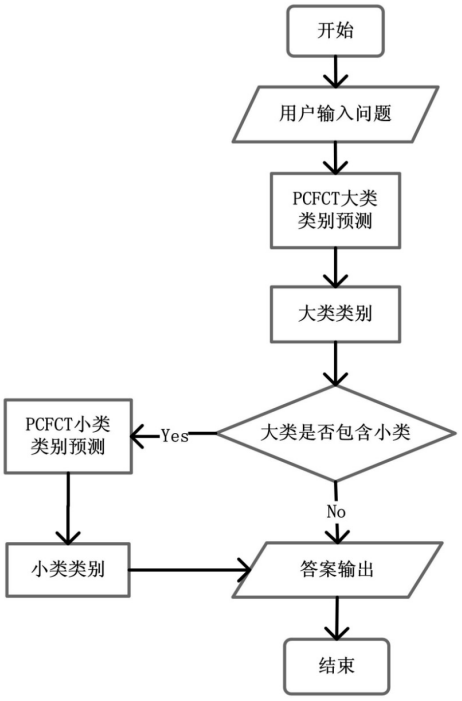

图 5　问答系统整体流程

2　实验与结果分析

2.1　实验设置与评价指标

PCFCT 方法和问答系统的实验环境采用 Pytorch2.2.1+cuda12.1 深度学习框架,编程语言使用 Python3.8,在 NVIDIA GeForce RTX 4060 Laptop GPU 显卡上进行计算,硬件配置是 12th Gen Intel(R) Core (TM) i7-12650H。在同一随机划分的数据集上,通过实验多次微调参数,选取效果最佳的一组参数作为正式的使用,设定参数优化器是 Adam, epoch 为 30, learing_rate 是 3e-5, max_seq_len 为 128, batch_size 是 8。

采用的评价指标是文本分类任务中常用的准确率(Acc)和 F1 值,Acc 表示正确分类的样本数占总样本数的比例,反映了模型整体的正确分类能力;F1 是精确率(Precision)和召回率(Recall)的调和平均值,在文中类别不平衡的数据集上更能反映模型的性能。

2.2　不同模板和模型对比

设计了八类不同的提示模板在不同的 PLM 上进行实验,选取三种 PLM(分别是 Bert、RoBerta、ALBert)进行对比实验,此实验部分是基于问题大类所设计的模板。Bert 模型是利用双向 Transformer 架构进行深度语义理解,通过预训练和微调两个步骤来提升各种

NLP 任务的性能;利用大规模的无标签文本数据,通过预训练来学习语言表示,用特定任务的标注数据进行模型的训练微调,使其能够适应这些任务的特定需求。RoBerta 模型对 Bert 模型进行了优化,引入了动态掩码技术,同时舍弃了 NSP 任务,预训练数据比 Bert 模型规模更大,RoBerta 模型通过一系列的优化策略和训练技巧来提升模型的性能。ALBert[30] 模型是 Bert 模型的精简版,对 Bert 模型的架构做了一些改变,在一定程度上缓解了大型预训练语言模型的资源消耗和训练时间过长的问题。表 3 列出了实验获取的表 2 所列不同模板和选取的 3 个 PLM 对应的 Acc、F1 值。

从表 3 可知,由于 ALBert 对模型架构做了改变,资源消耗和训练时间缩短了,但是分类效果没有另外两个模型好;RoBerta 比其他两个模型分类效果好,验证了 RoBerta 的改进提升了模型性能。由于需要分析[MASK]位置对分类结果的影响,所以计算了 P_1 ~ P_4 和 P_5 ~ P_8 对应三个 PLM 的 Acc 平均值,P_1 ~ P_4 对应的 Bert、RoBerta、ALBert 模型 Acc 平均值分别为 88.83%、90.60%、84.52%,P_5 ~ P_8 对应的 Bert、RoBerta、ALBert 模型 Acc 平均值分别为 90.51%、91.57%、85.47%,由此可以看出,模板中[MASK]位于句中,对上下文提示作用更明显,分类效果更好;P_1、P_5 模板长度较短,P_2、P_6 和 P_3、P_7 分别对应模板中包含任务关键词和领域关键词,均对最终的 Acc 和 F1 值有影响;综合对比数据,P_8 对应的 Acc 和 F1 值相对较高,该类模板中[MASK]位于句中且包含领域关键词和任务关键词,分析可知,模板提示越详细,对后续 PLM 影响越多,上下文理解能力越强,最终分类的效果就越好。

表 3 不同提示模板和模型的实验对比 %

模板	Bert		RoBerta		ALBert	
	Acc	F1 值	Acc	F1 值	Acc	F1 值
P_1	87.31	88.74	**89.83**	**90.16**	83.92	85.48
P_2	87.94	89.43	**90.51**	**91.22**	84.17	84.46
P_3	89.56	91.36	**90.85**	**92.53**	84.58	86.31
P_4	90.51	93.46	**91.19**	**93.74**	85.42	89.97
P_5	89.34	91.83	**90.34**	**92.49**	84.75	88.73
P_6	89.84	91.75	**91.53**	**92.13**	85.08	88.46
P_7	90.83	94.75	**91.86**	**95.87**	85.93	89.55
P_8	**92.03**	**96.43**	**92.54**	**97.94**	**86.10**	**90.61**

注:加粗字体是同行和同列中 Acc 和 F1 值数据的最高值。

2.3 不同分类方法对比

基于 2.2 节的实验结果,本节选择 RoBerta 模型作为 PCFCT 方法的 PLM 模型,因为对比模型中有基于 Bert 的文本分类,文中方法是基于 RoBerta 模型加入提示学习进行实验,所以对比可以看出提示学习消融前后的结果;并且选择提示模板 P_8:“文旅用户问题属于[MASK]类型的问题”进行对比分析。通过与 TextCNN、P-tuning 以及预训练语言模型 Bert、ERNIE[31] 基线方法进行对比实验,验证 PCFCT 方法的有效性。TextCNN 利用卷积神经网络(CNN)提取文本中的局部特征,并通过这些特征进行分类;P-tuning 将模板构建转化成参数优化问题,实现模板自动生成;Bert 通过预训练的双向 Transformer 模型来生成上下文相关的词向量,然后利用分类器进行分类;ERNIE 通过引入外部知识来增强表示能力,从而提升 NLP 能力。表 4 展示了各种基线方法与文中方法的实验对比结果,可以看出 PCFCT 方法对问题分类的效果较好;相比于基于 Bert 的文本分类方法,加入提示学习后准确率和 F1 值分别提升了 5.77 百分点和 5.78 百分点,提高了分类的准确率。

表 4 不同分类方法实验对比 %

方法	Acc	F1 值
TextCNN	77.73	85.63
P-tuning	84.34	88.57
ERNIE	85.79	91.30
BERT	86.77	92.16
PCFCT	**92.54**	**97.94**

2.4 实验结果分析

基于 2.2 节的实验结果,P_8 这一类的提示模板效果比其他模板好,所以问答系统构建过程中均基于 P_8 这类提示模板设计类别模板,例如:“这属于游览路线中的[MASK]类文旅问题”或者“该问题是文旅客服中的[MASK]的购票须知类”等,类别提示模板中包含类别名称且[MASK]位于句中、包含领域和任务关键词,该文对大类以及大类中的 15 个类别细化小类,基于 PCFCT 方法分别做了分类实验;同时,RoBerta 模型比其他 PLM 效果更优,所以小类训练所用 PLM 均为 RoBerta 模型;实验参数设置与大类训练设置相同;

评估指标使用准确率 Acc。本节基于大类和小类最终的结果计算出问答系统的准确率,如图 6 所示。

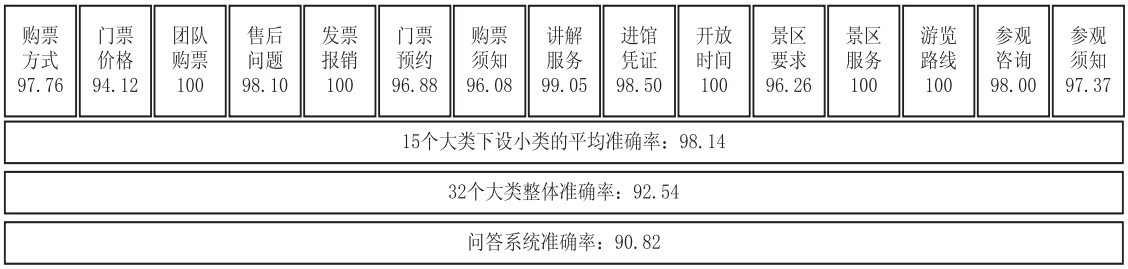

图 6　问答系统整体准确率(%)

从图 6 可以看出,问题小类类别分类实验准确率较高,但类别越多,分类效果相对越差,例如图中大类门票价格分了 11 个小类,其小类问题分类的准确率为 94.12%,是小类分类中准确率最低的,而小类划分只有 2 类的,如团队购票、景区服务等划分小类类别的准确率可以达到 100%,因此可见分类数量对 PFCTC 方法也有影响。为了评估问答系统的性能,本节计算了大类下设小类准确率的平均值作为小类分类的整体准确率为 98.14%;然后将大类最终的准确率 92.54% 与小类的整体准确率相乘计算得出问答系统的准确率为 90.82%。

3　结束语

该文提出了一种基于提示学习的小样本文旅客服问题分类方法(PCFCT),应用该方法构建了文旅客服问答原型系统。PCFCT 不需要大量的标注数据,在小样本数据的情况下,完成了文旅客服领域内的问题分类,模型充分利用了 PLM 的泛化能力,减少了对下游任务标签数据的依赖;把用户问题和带有[MASK]的提示模板拼接起来作为模型输入序列,对模型进行微调,通过标签预测和类别映射得到最终的分类结果。最后将 PCFCT 应用于文旅客服问答系统。实验表明,该方法提高了文旅客服问题分类的准确度。文中模板构造采用了原始的人工构造方法,在未来的工作中,将继续研究如何自动生成提示模板,应用于文旅领域中,使得提示学习的方法在文旅领域的可用性增强。

参考文献:

[1] 赵　芸,刘德喜,万常选,等. 检索式自动问答研究综述[J]. 计算机学报,2021,44(6):1214-1232.

[2] YU Wenhao,ZHU Chenguang,LI Zaitang,et al. A survey of knowledge-enhanced text generation[J]. ACM Computing Survey,2022,54(11s):1-38.

[3] YE Hongbin,LIU Tong,ZHANG Aijia,et al. Cognitive mirage:a review of hallucinations in large language models[DB/OL]. [2023-09-13]. https://doi. org/10. 48550/arXiv. 2309. 06794.

[4] 王静静,叶　鹰. 生成式 AI 及其 GPT 类技术应用对信息管理与传播的变革探析[J]. 中国图书馆学报,2023,49(6):41-50.

[5] 胡志强,李朋骏,王金龙,等. 基于 ChatGPT 增强和监督对比学习的政策工具归类研究[J]. 计算机工程与应用,2024,60(7):292-305.

[6] 闫　悦,郭晓然,王铁君,等. 问答系统研究综述[J]. 计算机系统应用,2023,32(8):1-18.

[7] HWANG S,LEE G. Conversational QA dataset generation with answer revision[C]//Proceeding of the 29th international conference on computational linguistics. Gyeongju:ICCL,2022:1636-1644.

[8] WANG Yaqing,YAO Quanming,KWOK J T,et al. Generalizing from a few examples:a survey on few-shot learning[J]. ACM Computing Surveys,2020,53(3):1-34.

[9] HE Kai,HUANG Yucheng,MAO Rui,et al. Virtual prompt pre-training for proto type-based few-shot relation extraction[J]. Expert Systems with Applications,2023,213:118927.

[10] BROWN T,MANN B,RYDER N,et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems,2020,33:1877-1901.

[11] 顾勋勋,刘建平,邢嘉璐,等. 文本分类中 Prompt Learning 方法研究综述[J]. 计算机工程与应用,2024,60(11):50-61.

[12] LIU Pengfei,YUAN Weizhe,FU Jinlan,et al. Pre-train,prompt,and predict:a systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys,2023,55(9):1-35.

[13] LIN Yun,WANG Jie. Research on text classification based on SVM-KNN[C]//2014 IEEE 5th international conference on software engineering and service science. NJ:IEEE,2014:842-844.

[14] LAI S,XU L,LIU K,et al. Recurrent convolutional neural networks for text classification[C]//Proceedings of the AAAI conference on artificial intelligence. Austin:AAAI,2015:2267-2273.

[15] CHEN Yahui. Convolutional neural network for sentence classification[D]. Waterloo:University of Waterloo,2015.

[16] LIU Yinhan,OTT M,GOYAL N,et al. Roberta:a robustly optimized bert pretraining approach[DB/OL]. [2019-07-26]. https://doi. org/10. 48550/arXiv. 1907. 11692.

[17] LU Zhibin, DU Pan, NIE J. VGCN - BERT: augmenting BERT with graph embedding for text classification[C]// Advances in information retrieval. [s. l.]: Springer, 2020: 369-382.

[18] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [DB/OL]. [2019-05-24]. https://doi. org/10. 48550/arXiv. 1810. 04805.

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

[20] ZHANG Z, WANG B. Prompt learning for news recommendation[C]//Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval. Taipei, China: SIGIR, 2023: 227-237.

[21] SONG Chengyu, CAI Fei, WANG Mengru, et al. TaxonPrompt: taxonomy-aware curriculum prompt learning for few-shot event classification[J]. Knowledge-Based Systems, 2023, 264: 110290.

[22] WALLACE E, FENG S, KANDPAL N, et al. Universal adversarial triggers for attacking and analyzing NLP[C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing. Hong Kong, China: EMNLP-IJCNLP, 2019: 2153-2162.

[23] SCHICK T, SCHÜTZE H. Exploiting cloze questions for few shot text classification and natural language inference[DB/OL]. [2021-01-25]. https://doi. org/10. 48550/arXiv. 2001. 07676.

[24] SHIN T, RAZEGHI Y, LOGAN I V R L, et al. Autoprompt: eliciting knowledge from language models with automatically generated prompts[DB/OL]. [2020-11-07]. https://doi. org/10. 48550/arXiv. 2010. 15980.

[25] ZHONG Zexuan, DAN Friedman, CHEN Danqi. Factual probing is [mask]: learning vs. learning to recall[C]//Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies. Online: ACL, 2021: 5017-5033.

[26] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation[C]//Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing. Online: ACL, 2021: 4582-4597.

[27] LIU Xiao, ZHENG Yanan, DU Zhengxiao, et al. GPT understands, too[DB/OL]. [2023-10-25]. https://doi. org/10. 48550/arXiv. 2103. 10385.

[28] LIU X, JI K, FU Y, et al. P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks [DB/OL]. [2022-03-20]. https://doi. org/10. 48550/arXiv. 2110. 07602.

[29] 王正佳,李 霏,姬东鸿,等. 基于多掩码与提示句向量融合分类的立场检测[J]. 计算机技术与发展,2023,33(12): 156-162.

[30] LAN Z, CHEN M, GOODMAN S, et al. Albert: a lite bert for self-supervised learning of language representations[DB/OL]. [2020-02-09]. https://doi. org/10. 48550/arXiv. 2010. 15980.

[31] SUN Yu, WANG Shuohuan, LI Yukun, et al. Ernie 2. 0: a continual pre-training framework for language understanding [C]//Proceedings of the AAAI conference on artificial intelligence. Austin: AAAI, 2020: 8968-8975.