大语言模型驱动下基于零样本语境联想的意图分类

陶汉卿,程玉虎,王雪松,王 军*
(中国矿业大学信息与控制工程学院,江苏徐州 221116)

摘 要: 意图分类是自然语言处理领域中的一项基础而关键的任务,其目标在于准确识别用户输入语句所表达的潜在意图,是对话系统、智能客服与人机交互等应用的重要技术支撑。近年来,基于深度学习的意图分类方法取得了显著进展,但其性能高度依赖大规模标注语料与稳定的领域分布,在实际应用中仍面临诸多挑战。尤其在短文本信息稀疏、标签语义抽象以及领域先验不足等低资源情境下,用户表达往往具有信息密度低、语义依赖隐含、表述方式多样等特点;同时,意图标签本身通常具有高度抽象性,不同标签之间语义边界模糊,现有模型难以仅凭文本内部的字面特征充分刻画深层语义与语境关联,进而制约了意图分类模型在低资源与跨场景条件下的泛化能力与鲁棒性。针对上述问题,本文从语义扩展与语境建模的角度出发,尝试突破传统监督学习对显式标注样本与表层字面特征的依赖。不同于将任务直接设定为零样本意图分类,本文在有监督学习框架下引入大语言模型的零样本语境联想能力,利用其蕴含的丰富世界知识与语义推理能力,扩展可学习的语义空间,从而弥补文本信息稀疏与标签语义不足所带来的建模缺陷。基于这一思路,本文提出一种基于大语言模型的零样本语境联想模型(LLM-based Zero-shot Context Association Model,L-ZCAM)。该模型通过构造结构化提示词,引导大语言模型从联想意图与标签定义两个互补视角生成与输入语句相关的补充性语境语义信息,实现文本内部特征与文本外部知识的联合挖掘,并对意图标签的语义内涵进行显式增强。在模型结构设计上,L-ZCAM 采用多路特征编码与交叉注意力机制,对原始文本特征、联想语义特征及标签语义特征进行深度交互建模;同时,引入约束引导的联合损失函数,对联想语义与标签语义之间的一致性进行约束,以缓解语义噪声带来的干扰,实现文本内外信息的有效对齐。通过上述设计,L-ZCAM 能够更好地感知多义模糊、标签抽象以及表达多样等复杂语境下的语义关联关系,从而提升意图判别的准确性与稳定性。实验结果表明,在 CLINC150、Banking77 和 HWU64 三个公开数据集上,L-ZCAM 的宏平均 F1 分数分别较当前最新方法提升 2.25%、1.28% 和 1.29%,在不同任务场景下具有更强的泛化能力与鲁棒性。

关键词: 大语言模型;意图分类;零样本语境联想;语义扩展;特征生成;交叉注意力
基金项目: 中国博士后科学基金(No.2024M753519,No.GZC20241922);江苏省卓越博士后资助项目(No.2024ZB721)
中图分类号: TP181;TP183 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0219-15
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250863

An Intent Classification Method Based on Zero-Shot Context Association Driven by Large Language Models

TAO Hanqing, CHENG Yuhu, WANG Xuesong, WANG Jun*
(*School of Information and Control Engineering*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116, *China*)

Abstract: Intent classification is a fundamental and critical task in natural language processing, aiming to accurately identify the underlying intentions expressed in user utterances. It serves as an essential technical foundation for dialogue systems, intelligent customer service, and human-computer interaction. In recent years, deep learning-based approaches have achieved remarkable progress in intent classification; however, their performance heavily relies on large-scale annotated corpora and stable domain distributions, which poses significant challenges in real-world applications. In low-resource scenarios characterized by sparse short-text information, abstract label semantics, and insufficient domain prior knowledge, user expressions often exhibit low information density, implicit semantic dependencies, and diverse surface forms. Meanwhile, intent labels are typically highly abstract with blurred semantic boundaries, making it difficult for existing models to capture deep semantic representations and contextual associations solely from literal textual features. These issues severely limit the generalization ability and robustness of intent classification models under low-resource and cross-domain settings. To address these challenges, this paper explores intent classification from the perspective of semantic expansion and contextual modeling, aiming to reduce the reliance of traditional supervised learning methods on explicit annotations and shallow lexi-

收稿日期:2025-09-28;录用日期:2026-01-06;责任编辑:覃怀银
*通信作者:王军

cal features. Unlike approaches that directly formulate the task as zero-shot intent classification, we introduce the zero-shot contextual association capability of large language models into a supervised learning framework. By leveraging the rich world knowledge and semantic reasoning ability encoded in LLMs, the proposed approach expands the learnable semantic space, thereby alleviating the modeling limitations caused by sparse textual information and insufficient label semantics. Based on this idea, we propose an LLM-based zero-shot context association model (L-ZCAM). The model constructs structured prompts to guide LLMs to generate complementary contextual semantic information related to the input utterance from two complementary perspectives: associative intents and label definitions. This design enables joint mining of in-text features and out-of-text knowledge while explicitly enhancing label semantics. From a structural perspective, L-ZCAM adopts multi-branch feature encoders and a cross-attention mechanism to deeply model the interactions among original textual features, associative semantic features, and label semantic features. In addition, a constraint-guided joint loss function is introduced to enforce semantic consistency between associative semantics and label semantics, mitigating the impact of semantic noise and achieving effective alignment between internal and external information. Through these designs, L-ZCAM is able to better capture semantic associations under complex contexts involving polysemy, abstract labels, and diverse expressions, thereby improving the accuracy and stability of intent prediction. Experimental results on three public datasets, i.e., CLINC150, Banking77, and HWU64, demonstrate that L-ZCAM outperforms state-of-the-art methods by 2.25%, 1.28%, and 1.29% in terms of macro-averaged F1 score, respectively, exhibiting stronger generalization ability and robustness across different task scenarios.

Keywords: large language model; intent classification; zero-shot context association; semantic expansion; feature generation; cross attention

Foundation Item(s): China Postdoctoral Science Foundation (No.2024M753519, No.GZC20241922); Jiangsu Funding Program for Excellent Postdoctoral Talent (No.2024ZB721)

0 引言

在自然语言处理领域中,意图分类(intent classification)任务作为构建智能问答系统、人机对话系统、信息检索系统等高层语言理解模块的基础,其核心目标是从用户输入的自然语言文本中识别其潜在的语义意图[1]。然而,在常见的对话系统、语音助手或客服机器人中,用户输入的查询往往具有高度的表达自由性和不确定性,甚至带有隐喻或多义的方式表达需求。此类表达方式具有极高的信息熵,缺乏稳定的字面特征,且需要依赖大量文本字面语义之外的语境信息和先验知识储备来实现正确理解[2]。如图1所示,以航空客服场景为例,当用户提出“我的航班因台风取消,能否改签至后天同一时段并保留升舱资格?”这一复杂请求时,传统模型往往依赖关键词匹配,若未及时在数据库中添加“台风改签”或“升舱资格保留”等新涌现的意图标签数据,往往会将新业务下收到的用户请求错误归类为过时、通用的“航班改签”“航班升舱”等服务类别,进而在无法识别上下文逻辑关系的前提下,对“改签”和“升舱”分别响应模板化内容,未能正确理解用户的真正意图和相应业务的分类,最终导致后续业务流程处理失败。

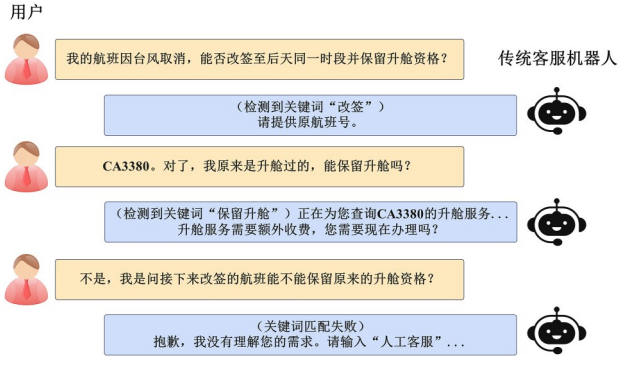

图1 传统意图分类客服系统示例

Figure 1 Example of a traditional intent classification service system

这一案例暴露出当前智能化人机交互系统在处理语义组合、对话上下文跟踪以及意图推理等方面的不足,可以为此类用户查询对话文本归纳出如下三个典型特点:一是信息稀疏,即短文本中词汇数量有限,缺乏足够的上下文结构支撑语义推理;二是语义抽象,平台业务的标签本身往往是人为概括而来的语义范畴,其语言外显形式与文本之间的联系较为隐含;三是先验缺失,在开放领域的应用中,模型通常缺乏对特定用户背景、任务目标的知识理解。这三重挑战导致传统基于字面特征或静态语义表示的意图分类方法在零样本情境下表现严重受限。为了缓解这一困境,近年来大量研究开始尝试引入预训练语言模型(pretrained language model),借助其通用语义表征能力进行下游任务迁移[3]。然而,当前主流方法多数仍停留于“字面建模”层面,还只是停留在将输入文本通过大模型编码表征后直接送入分类器的阶段,未能深入考量文本背后蕴含的先验知识、或是挖掘人

类语言理解过程中联想机制与标签建构之间的语义桥梁,模型在面对文本表达多样性、标签语义模糊性时,仍难以实现精准的语义对齐与泛化识别[4]。

实际上,用户意图往往复杂多样,常常暗含多种领域知识,并可能通过隐晦表达、上下文暗示、文化或领域特定的交流方式体现出来,甚至会掺杂各种口语,远远超出了传统分类模型所能处理的范畴。因此,在人工智能与人机交互领域,意图分类作为对话系统、智能客服等应用的核心技术,长期面临标注数据依赖性强与领域泛化能力弱的双重挑战[5]。传统监督学习方法虽然在限定领域内表现优异,但其性能高度依赖于大规模标注数据和关键词匹配过程,难以灵活、有效应对现实场景中不断涌现的新领域、新意图需求[6]。与此同时,完全依赖零样本意图分类去处理未见类别,往往又因缺乏稳定的类别原型而难以保证判别可靠性;更为普遍的情形是样本有限、标签语义抽象、上下文信息不足,导致模型难以充分挖掘潜在语义与跨场景共性特征、在低资源场景下的泛化能力较弱。因此,在不将问题简单设定为零样本分类的前提下,如何有效借助外部知识源增强模型的语境联想与理解能力,成为提升有监督模型泛化性能的关键思路。

从认知心理学的视角来看,人类在理解语言时并非完全依赖文本的字面信息,而是会主动调动自身的语言经验与世界知识,构建以关键词为中心的概念网络,从而形成一种“以词激发联想”的语境加工方式[7]。这种联想机制在面对篇幅短、线索少的文本时尤为关键,因为它体现出人类在信息不完备条件下,如何通过知识迁移和语义补全实现对语言的深层次理解。同时,在分类判断的过程中,人脑也并非将标签看作孤立符号,而是将其作为一个语义范畴的代表,背后往往伴随着一整套语言经验与范畴知识[8]。这意味着,标签不仅是模型学习的目标输出,也应是语义理解过程中的参与者,具有解释性和指导性。因此,要想实现智能系统对复杂意图的有效识别,必须从两个关键方向着手:一是引入类人联想能力,通过上下文语义触发多层次语境推理,从而增强模型对文本语义的泛化与联想建构能力;二是赋予标签语义定义和结构化含义,使模型能够利用标签所蕴含的外部语义信息,在推理过程中形成自上而下的概念指引。

随着当下以 ChatGPT 为代表的大语言模型(Large Language Models,LLMs)的快速发展[9-11],上面的设想逐渐变得可行,学术界和工业界开始重新审视意图分类任务的建模方式[12]。这些大语言模型在海量语料上进行预训练,具备深层语义表示、上下文联想能力以及生成式推理能力,能够在零样本或少样本条件下生成重要的先验知识内容,进而辅助完成多种自然语言处理领域任务,被广泛认为是零样本语义理解的重要突破[13]。已有研究表明,LLMs 在开放领域对话生成、人类指令跟随以及含糊指令推理等方面表现出令人惊艳的能力,甚至能在没有明确标签监督的前提下,对用户意图进行合理猜测[14]。而目前大多数意图分类工作仍集中于基于微调 Transformer 系列模型的传统分类模式[15],对如何系统性地激发语言模型在上下文理解和语义联想方面的潜力,尤其在意图分类中进行生成式推理与多轮理解,尚缺乏深入探索。

基于以上认知启发与当下的任务挑战,本研究提出了一种基于大语言模型的零样本语境联想模型(LLM-based Zero-shot Context Association Model, L-ZCAM)用于文本意图分类。该模型融合了大语言模型强大的知识生成能力和语境联想能力,旨在模拟人类语言理解中“文本—联想—标签”三层认知路径,在零样本语境下构建对用户意图的深度挖掘与建模。整体而言,L-ZCAM 模型包括输入层、表征层和分类层三大模块,其中的关键设计在于通过结构化提示词分别引导大语言模型生成与文本相关的联想意图原型与对应的摘要定义,同时构建对标签集合的语义定义,从而在编码层以三方语义表示为基础,通过交叉注意力机制实现三者之间的深层语义对齐与信息交互。这一设计不仅弥补了短文本表达不足的问题,也赋予模型以标签语义主导下的多层次推理能力,从而在语义空间实现更为稳健的意图聚合与分类判断。

1　相关工作

意图分类作为自然语言理解的核心任务,其方法演进始终与标注数据规模、领域泛化需求及上下文建模能力紧密相关。早期研究主要依赖基于规则与模板的方法,通过人工定义意图关键词与句式结构实现有限场景的意图分类[16]。例如,航空订票领域的规则系统可通过预设“改签”“退票”等触发词匹配用户请求。然而,此类方法严重受限于规则覆盖范围,难以处理语言表达的多样性与歧义性。随着深度学习的兴起,监督学习模型逐渐成为主流,研究者先后尝试使用 RNN、CNN 及 Transformer 架构从标注数据中学习意图特征表示。其中,基于 BERT(Bidirectional Encoder Representations from Transformers)的意图分类器通过预训练-微调范式在特定领域达到较好的准确率[17]。然而,这些模型本质上仍是封闭世界假设下的分类器,其性能高度依赖目标领域标注数据的质量与规模,当测试意图类别未出现在训练集中时,传统 BERT 模型的准确率将显著下降,暴露出监督学习方

法在零样本场景下的根本性缺陷。

为降低对标注数据的依赖，研究者开始探索零样本学习（zero-shot learning）在意图分类中的应用。零样本学习是一种突破性的学习范式，旨在让机器学习模型能够像人类一样，利用已有的知识和理解去推理和识别从未直接见过的事物，其核心在于利用语义信息（属性、词向量、描述等）作为连接已知世界（训练类）和未知世界（测试类）的桥梁，并通过映射或生成模型实现知识的迁移[18]。尽管面临着领域漂移、枢纽点等挑战，零样本学习代表了构建更具泛化能力、适应新环境能力的人工智能的重要一步。其核心思想是通过语义属性迁移或跨模态对齐实现语义迁移与语境外延扩展。早期零样本方法主要采用属性映射策略，例如将意图描述映射到共享语义空间或通过知识图谱构建意图间的层级关系[19]。进一步地，语义嵌入迁移方法通过对比学习将判别标签与用户语句编码至同一向量空间，显著提升了零样本泛化能力[20]。Sentence-BERT[21]则通过孪生网络结构计算语句与意图描述的相似度，在零样本意图分类任务中取得突破性进展。但是，此类方法仍存在两大局限：其一，意图原型通常由静态的标签描述生成，无法捕捉对话语境中动态演变的语义线索；其二，语义相似度计算忽略了语句结构与逻辑推理的重要性，导致对复合意图的识别能力不足。因此，本文并未将问题设定为零样本意图分类，而是在零样本思想下关注更为常见的低资源情境，并在有监督框架下引入大语言模型的零样本语境联想能力，以动态扩展标签语义与上下文关联，从另一条路径提升模型的泛化与鲁棒性。

幸运的是，当下大语言模型的崛起为低资源场景下依赖外部知识源辅助的意图分类提供了新的技术路径。GPT-4[9]、PaLM[10]以及LLaMA[11]等大语言模型通过海量文本预训练获得的上下文推理能力，可在极少甚至零标注样本下完成复杂语言任务。研究者尝试通过不基于任何数据样本的零样本提示策略（zero-shot prompting）来激活LLMs的隐式知识，将意图分类转化为文本生成任务[22]。然而，直接将LLMs用于零样本意图分类面临三重挑战：首先，生成式模型的输出具有不可控性，可能产生偏离预设标签体系的意图描述；其次，LLMs的推理过程缺乏可解释性，难以定位语境关联的关键证据；最后，模型的计算开销限制了其在实时系统中的应用。针对这些问题，近期工作提出轻量化适配策略，例如使用LoRA[23]对LLMs进行参数高效微调，或通过知识蒸馏将LLMs能力迁移至小型分类器[24]。IntentGPT[25]通过提示模板引导LLM生成意图相关的逻辑链，再使用注意力机制提取关键推理步骤作为分类依据。进一步地，MetaTCN采用了元学习框架，利用三重对比学习同时优化双向知识嵌入，并采用动态变化率采样策略对具有挑战性的样本进行优先排序[26]。DLNR[27]则是一种基于大语言模型的小样本对话意图分类方法，采用动态标签精化技术，具备动态与上下文感知能力。相较于静态方法，它可根据不同查询动态调整标签，从而更精准地捕捉语境化语义关系，并保持原始意图之间的语义关联。最近，一种多语义对比学习模型（Multi-Semantic Contrastive Learning Method，MSCLM）被提出，它采用交叉注意机制，通过基于大语言模型提示的对比学习来对齐字面意义和隐喻意义，从而缓解字面语义与隐喻语义不一致的问题[28]。以上这些方法虽在语义关联和跨领域任务方面展现了较好的性能和基于大模型的开发潜力，但总体来说，它们性能仍受限于LLM的语境建模粒度——现有研究多将整个用户交互语句作为连续文本输入，未能有效引入智能系统所涉及的业务领域、场景偏好等动态要素，往往忽视了不同层次的语境依赖，从而启发了本文后续开展的探索和研究内容。

2 问题定义

意图分类通常被形式化为一个多分类问题，因此这里首先给出所研究内容的问题定义。在本文中，给定一个用户输入的查询文本T，目标是从一个预定义的意图标签集合中，通过有监督的模型训练过程，为其分配一个最匹配的意图标签$y \in Y$，其中$Y=\{y_1, y_2, \cdots, y_K\}$，$K$为数据集的标签数量。形式上，希望学习一个映射函数$f: T \to Y$，该函数能够捕捉用户查询背后的语义，并准确预测其意图。

然而，在真实应用中，意图标签分布呈现长尾特性，许多标签具有领域依赖性，且缺乏足够的人工标注数据，尤其是新意图往往没有任何标注样本，与用户输入的查询文本没有任何实质意义上的交互和语义联系。为解决这一问题，本文在任务设定上与零样本分类相区分开，研究重心在于探索如何在没有任何参照语境样本的情况下引导大语言模型进行零样本语境联想，为有监督意图分类任务生成、引入关键的用户查询背后可能涉及的先验语境信息，实现特征增广，进而丰富用户输入查询文本的表示。为此，可以将问题形式化定义如下：

$$f(T)=\arg\max_{y \in Y} P\left(y|T, A(T), D(Y), t_{\text{dataset}}\right) \quad (1)$$

其中，$P(\cdot|\cdot)$代表条件概率计算；$A(T)$代表大语言模型基于用户查询输入文本T生成的零样本语境联想特征；$D(Y)$代表大语言模型基于意图标签集合Y生成的标签定义；t_{dataset}则为数据集的任务描述。

3　L-ZCAM模型

在阐述模型的具体结构与实现细节之前,有必要首先明确本研究的核心设计理念与理论依据。

其一,从人类的意图本质来看,大脑执行阅读与语义理解任务时并非仅依赖文本表层线索,而是会主动调动先验知识与概念网络,通过联想式加工完成深层语义推理[29]。在短文本、信息稀疏或表达模糊的情境中,大脑则倾向于借助长期积累的语言经验和世界知识,实现跨语境、跨文本的语义泛化与意义推断[30]。这种自然的联想认知过程构成了本研究提出零样本语境联想式意图建模的理论基础。

其二,从人类对事物分类的本质来看,意图分类并非简单的类别映射过程,而是要理解文本内容与标签语义之间的隐含对应关系[31]。传统的意图分类方法对于标签进行one-hot的形式化"哑元编码"方式虽便于计算,但实际上标签被视为无语义的符号化标记,模型无法感知标签所蕴含的概念属性或语义指向[32]。若将标签语义或结构化语义显式引入模型,使标签从"哑元编码"转化为"语义参照物",即可强化模型对文本与类别之间语义关联的建模能力,这构成了本研究采用"文本—联想—标签"约束式对比学习建模的理论基础。

因此,基于以上两方面的认知出发点,为有效解决实际智能系统接收到用户查询文本信息熵高、字面特征稀疏、标签语义抽象以及表达形式多样等问题,本文设计了一种基于大语言模型的零样本语境联想模型,旨在利用大模型的知识生成能力和推理联想能力,从多层次语义视角进行意图建模与分类推断。L-ZCAM的整体模型结构如图2所示,主要包括:输入层、表征层、分类层,每个部分的详细内容将在后文依次展开描述。

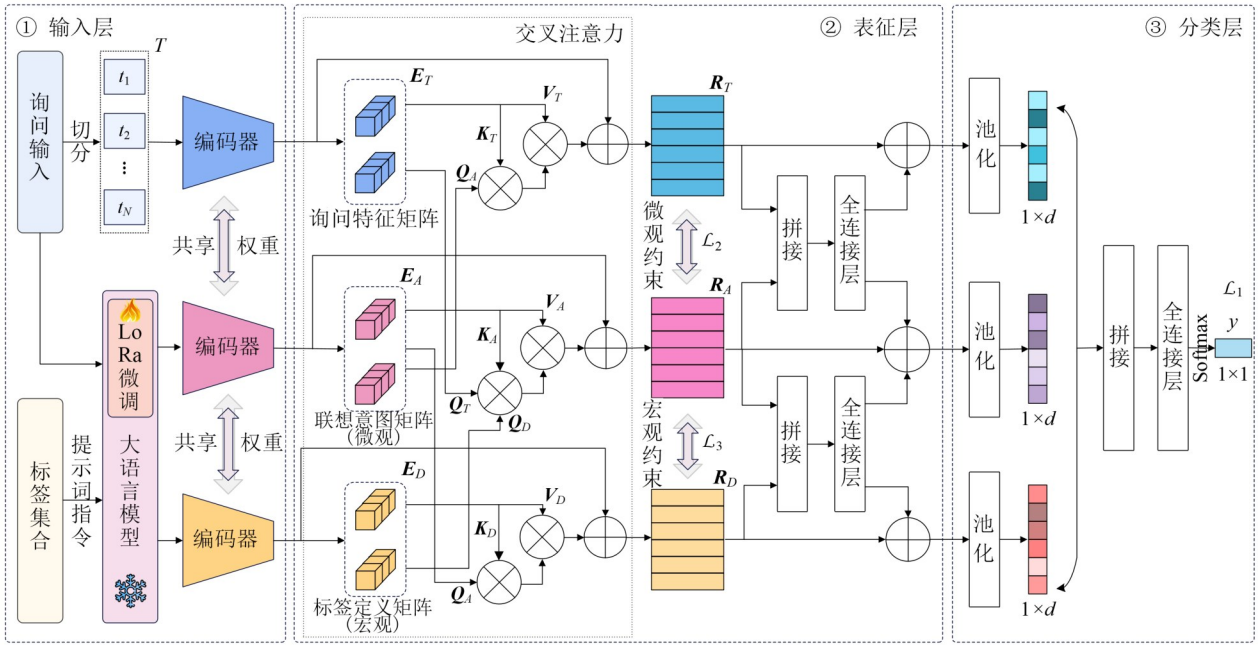

图2　L-ZCAM模型整体示意图

Figure 2　The overall architecture of L-ZCAM model

3.1　输入层

如图2所示,输入层的主要目的是整合用户查询输入、标签集合内容,构造结构化提示词输入给LLM以生成联想意图特征和标签定义特征。在进行后续如图3所示的引导大语言模型进行特征生成的细化流程之前,对于用户的询问输入文本T,需要将其按分词器(tokenizer)进行切分,得到N个单元token序列:

$$T=[t_1,t_2,\cdots,t_N] \tag{2}$$

其中,N为询问输入文本T切分得到的token数量。

首先,在数据集的宏观层面上,结合标签集合Y及数据集任务描述t_{dataset},可以构造标签定义提示词p^l,引导大语言模型为每个常被忽视的标签文本生成摘要性的一句话定义描述,进而获得数据集相应的标签定义集合:

$$D(T)=\text{LLM}(Y,p^l,t_{\text{dataset}}) \tag{3}$$

然后,为了实现零样本联想意图特征的构建,本文提出基于原型蒸馏和语义对齐的两阶段策略。

第一阶段,先通过设计结构化提示词p^a以引导大语言模型生成联想意图集合:

$$A(T)=[a_1,a_2,\cdots,a_M]=\bigcup_{i=1}^{M}\text{LLM}(T,p_i^a) \tag{4}$$

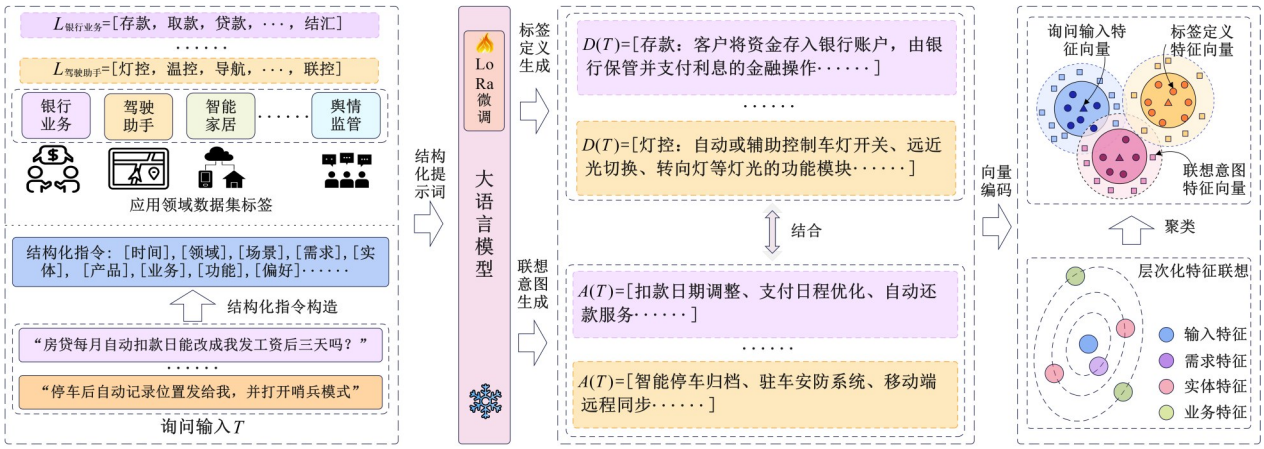

图3 引导大语言模型进行特征生成的示意图

Figure 3 Schematic diagram of orienting the large language model for feature generation

其中，超参数M为构建结构化指令时需要生成意图原型的层次数量（如：需求特征、实体特征、业务特征等）；p_i^a为指定了需求、实体、业务等细分引导的指令。在真实任务场景中，用户所提供的查询指令对应到智能系统或智能设备的执行顺序为：用户需求→涉及的实体→业务动作/流程。因此这里“层次化特征”指的是将用户输入携带的语义信息按照从抽象到具体、从目标到操作、从意图到系统实现路径的结构逐级展开的特征表达方式。

进一步，将得到的联想意图原型a_i依次输入大语言模型，像生成标签定义一样，提取a_i的摘要性描述ε_i：

$$\varepsilon_i=\mathrm{LLM}(a_i) \tag{5}$$

比如对用户输入的查询文本特征所生成的联想意图a_i：“节能模式自动化”，采用大语言模型为此生成的摘要性描述为ε_i：“用户希望系统在特定时间或条件下自动调整设备状态以降低能耗，通常涉及温度、照明等多设备联动”。

为了整合联想语境特征进而后续与用户询问输入文本、标签定义相交互，这里需要将多层次的联想意图原型a_i、联想意图摘要性描述ε_i逐一编码为向量：

$$\boldsymbol{e}_{a_i},\boldsymbol{e}_{\varepsilon_i}=\mathrm{BERT}(a_i,\varepsilon_i) \tag{6}$$

其中，i代表M个联想层次的索引。

由于需要确保大语言模型生成的联想语境特征与用户输入的查询文本是相关且匹配的，进一步设计了一种基于相似度的对齐策略进行关系界定。具体来说，首先计算每个结构化指令下生成的联想意图原型向量$\boldsymbol{e}_{a_i}(1\leqslant i\leqslant M)$与联想意图原型摘要性描述$\boldsymbol{e}_{\varepsilon_j}(1\leqslant j\leqslant M)$之间的相似度得分：

$$s_{ij}=\frac{\boldsymbol{e}_{a_i}\cdot\boldsymbol{e}_{\varepsilon_j}}{|\boldsymbol{e}_{a_i}||\boldsymbol{e}_{\varepsilon_j}|} \tag{7}$$

这里，根据意图向量之间的相似度计算、按照联想层次数量聚类为M个簇，便可以实现零样本语境联想情况下不同意图向量之间的距离衡量和界定，确保结构化提示词得到的生成内容是层次化的。

在基于原型蒸馏和语义对齐策略的第二阶段，将所得到的不同层次的联想意图原型摘要性描述$\boldsymbol{e}_{\varepsilon_i}$进行整合，并基于自注意力机制汇集成一个整体向量：

$$\boldsymbol{e}_A=\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{e}_{\varepsilon_i} \tag{8}$$

$$\alpha_i=\mathrm{Softmax}\left(\frac{\boldsymbol{e}_{\varepsilon_i}\cdot\boldsymbol{e}_A^{\mathrm{T}}}{\sqrt{d}}\right) \tag{9}$$

$$\boldsymbol{E}_A=\sum_{i=1}^{M}\alpha_i\boldsymbol{e}_{\varepsilon_i} \tag{10}$$

其中，$\boldsymbol{E}_A$为聚合得到的整体语境联想意图向量。

至此，通过设计的结构化提示词，L-ZCAM模型能够分别获取两类关键的语义信息：其一是从宏观层面提取的标签语言定义，作为任务语义的结构化先验；其二是围绕具体用户查询自动生成的微观语义联想，用以模拟人类在面对模糊或信息不足文本时所进行的意义扩展与概念激活。前者为模型提供类别级的语义框架和解释性边界，后者则作为与输入文本紧密相关的语境补全机制，为隐含意图的捕捉提供动态且细粒度的语义支撑。两者在大语言模型的生成过程中形成相互约束、相互促进的目标导向性，使得生成的联想内容在保持多样性的同时兼具语义一致性与任务相关性。

接下来，采用参数共享的BERT编码器，对用户的询问输入文本T与标签定义集合$D(T)$进行统一的形式化编码：

$$\boldsymbol{E}_T,\boldsymbol{E}_D=\mathrm{BERT}\left(T,D(T)\right) \tag{11}$$

其中,$\boldsymbol{E}_T,\boldsymbol{E}_D\in\mathbb{R}^{1\times d}$;$d$ 为编码器 BERT 的向量输出维度。

3.2 表征层

表征层旨在在连续的空间中对从输入层得到的三类编码向量进行深度的语义建模和挖掘。

首先,考虑到用户的询问输入文本是获取语境联想意图 $A(T)$ 的重要依据,而标签定义集合 $D(T)$ 又是引导语境联想意图 $A(T)$ 的监督内容,本文设计了针对性的交叉注意力机制,用以增强三种表示向量之间的语义交互。具体来说,交叉注意力机制计算过程的核心是交叉采用彼此作为自己的查询(query)向量,键(key)向量和值(value)向量则为自身的计算方式。相应地,以询问输入特征为出发点的交叉注意力机制权重矩阵计算过程如下:

$$\boldsymbol{Q}_T=\boldsymbol{E}_A\boldsymbol{W}_Q,\boldsymbol{K}_T=\boldsymbol{E}_T\boldsymbol{W}_K,\boldsymbol{V}_T=\boldsymbol{E}_T\boldsymbol{W}_V \tag{12}$$

$$\mathbf{Attn}_T=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_T\boldsymbol{K}_T^{\mathrm{T}}}{\sqrt{d}}\right) \tag{13}$$

$$\boldsymbol{H}_T=\mathbf{Attn}_T\boldsymbol{V}_T \tag{14}$$

进一步地,为了动态融合全局的特征信息与局部的实体特征,设计了一种实体状态门控机制如下:

$$\boldsymbol{G}_e=\sigma(\boldsymbol{H}_T\boldsymbol{W}_g+\boldsymbol{b}_g) \tag{15}$$

$$\boldsymbol{R}_T=\boldsymbol{G}_e\odot\boldsymbol{H}_T+(1-\boldsymbol{G}_e)\odot\boldsymbol{E}_T \tag{16}$$

其中,σ 为 Sigmoid 函数;$\odot$ 表示逐元素乘运算。

为描述方便和简洁,按照查询向量、键向量和值向量的顺序,式(12)~(16)可形式化为

$$\boldsymbol{R}_T=\mathrm{CrossAttn}(\boldsymbol{E}_A,\boldsymbol{E}_T,\boldsymbol{E}_T) \tag{17}$$

相应地,作为对偶设计,联想意图与标签定义之间的交叉注意力以及标签定义与联想意图之间的交叉注意力计算分别为

$$\boldsymbol{R}_A=\mathrm{CrossAttn}(\boldsymbol{E}_T,\boldsymbol{E}_A,\boldsymbol{E}_A) \tag{18}$$

$$\boldsymbol{R}_T=\mathrm{CrossAttn}(\boldsymbol{E}_A,\boldsymbol{E}_T,\boldsymbol{E}_T) \tag{19}$$

上述注意力机制均基于多头结构实现,用以建模不同语义表示之间的联想与对应关系。

随后,将各向量进行拼接并引入残差连接构建更深层表示:

$$\begin{cases}\boldsymbol{R}'_T=\mathrm{Res}\left(\mathrm{FC}\left([\boldsymbol{R}_T,\boldsymbol{R}_A]\right)\right)+\boldsymbol{R}_T\\ \boldsymbol{R}'_A=\mathrm{Res}\left(\mathrm{FC}\left([\boldsymbol{R}_A,\boldsymbol{R}_T]\right)\right)+\boldsymbol{R}_A\\ \boldsymbol{R}'_D=\mathrm{Res}\left(\mathrm{FC}\left([\boldsymbol{R}_D,\boldsymbol{R}_A]\right)\right)+\boldsymbol{R}_D\end{cases} \tag{20}$$

其中,[·,·]表示向量拼接操作;FC(·)表示全连接变换;Res(·)表示残差连接操作。这种相互迭代式增强方法使得用户输入的查询特征不仅能从自身表征中学习,还能在控制梯度消失问题的同时从联想推理与标签语义两端进行信息融合。

3.3 分类层

分类层旨在对上述表征层获得的语义向量进行压缩、融合与分类。首先,对三路输出 $\boldsymbol{R}'_T$、$\boldsymbol{R}'_A$、$\boldsymbol{R}'_D$ 分别进行平均池化操作 Pool(·),得到三个固定维度的一维表示:

$$\boldsymbol{R}''_T,\boldsymbol{R}''_A,\boldsymbol{R}''_D=\mathrm{Pool}(\boldsymbol{R}'_T,\boldsymbol{R}'_A,\boldsymbol{R}'_D) \tag{21}$$

其中,$\boldsymbol{R}''_T,\boldsymbol{R}''_A,\boldsymbol{R}''_D\in\mathbb{R}^{1\times d}$。

随后,将三者拼接为一个整体表征向量:

$$\boldsymbol{R}_{\mathrm{final}}=[\boldsymbol{R}''_T,\boldsymbol{R}''_A,\boldsymbol{R}''_D]\in\mathbb{R}^{1\times 3d} \tag{22}$$

最终的分类标签预测通过一层全连接层与 Softmax 函数实现:

$$y=\mathrm{Softmax}\left(\mathrm{FC}(\boldsymbol{R}_{\mathrm{final}})\right) \tag{23}$$

至此便实现了用户输入、联想意图与标签语义三者在语义空间的统一匹配,从而在零样本条件下实现鲁棒的意图分类。

3.4 损失函数设计

在上面的表征层设计中,交叉注意力机制具有关键的信息筛选和加权作用。但深入分析之后可以发现,输入文本 T 属于字面信息,而经过大语言模型生成的联想意图 A 和标签定义 D 均属于文本外先验信息。实际上,这两种文本外先验信息在性质上也存在着较大差异,即:联想意图需要依据具体的用户输入查询进行生成,而标签语义属于数据集的宏观信息,不依赖于具体的输入文本内容。也就是说,依赖于每个输入文本内部特征生成的联想意图是随着 L-ZCAM 模型训练过程动态生成并不断调整优化的,在丰富输入文本 T 表征的作用上起着直接作用;而不依赖于每个输入文本内部特征的标签定义是依据数据集信息,在 L-ZCAM 模型开始训练之前一次性生成获取的,属于宏观信息,对于动态生成的联想意图特征则起着监督和限制作用,确保联想意图特征是同时受限于输入内容和标签领域的约束。

因此,为有效整合和平衡输入的文本查询字面信息、联想信息以及标签定义信息,本文设计了一个联合损失函数:

$$\mathcal{L}=\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3 \tag{24}$$

其中,$\mathcal{L}_1$ 为训练过程中占据主体作用的交叉熵分类损失:

$$\mathcal{L}_1=-\sum_{k=1}^{K}y_k\log\hat{y}_k \tag{25}$$

其中,y_k 为真实标签的独热编码;$\hat{y}_k$ 为模型对第 k 类的预测概率。经过这种损失函数设计,可以在保持分类性能的同时,通过引导模型在语义空间对齐联想路径与标签语义定义,从而提升对复杂、隐含、多义表达

中意图的识别能力。

$\mathcal{L}_2$是经过交叉注意力机制计算得到的询问输入文本向量表征与联想意图向量表征$\boldsymbol{R}_A$之间的KL散度损失，用于对齐$\boldsymbol{R}_T$与$\boldsymbol{R}_A$的分布，增强模型以询问输入文本为依据进行联想的合理性：

$$\mathcal{L}_2=\mathrm{KL}(\boldsymbol{R}_T\|\boldsymbol{R}_A) \tag{26}$$

$\mathcal{L}_3$则是经过交叉注意力机制计算得到的联想意图向量表征$\boldsymbol{R}_A$与标签定义向量表征$\boldsymbol{R}_D$之间的KL散度损失，用于对齐$\boldsymbol{R}_A$与$\boldsymbol{R}_D$的分布，确保语境联想意图处于标签定义的监督引导：

$$\mathcal{L}_3=\mathrm{KL}(\boldsymbol{R}_A\|\boldsymbol{R}_D) \tag{27}$$

需要指出的是，在L-ZCAM模型搭建过程中，为了实现预训练大语言模型在意图分类领域的高效迁移，固定住大模型的参数不变，并采用了LoRA低秩策略注入了一定的可训练和微调的参数：

$$\boldsymbol{W}=\boldsymbol{W}_0+\boldsymbol{BA} \tag{28}$$

其中，$\boldsymbol{W}_0\in\mathbb{R}^{d\times d}$为原始权重；$\boldsymbol{B}\in\mathbb{R}^{d\times r}$，$\boldsymbol{A}\in\mathbb{R}^{r\times d}$为低秩矩阵，仅训练$\boldsymbol{B}$和$\boldsymbol{A}$以保留大语言模型的通用能力。

4 实验与分析

4.1 数据集描述

为全面评估L-ZCAM模型在不同类型的意图分类任务中的泛化能力与鲁棒性，本文基于自然语言处理平台Huggingface选取了三个具有代表性、应用场景差异明显的公开意图分类数据集进行实验验证，相应的统计信息如表1所示。

表1 数据集的统计信息

Table 1 Dataset statistics

数据集	CLINC150	Banking77	HWU64
样本总数量	23 700	13 083	25 716
训练集大小	17 775	10 003	19 287
测试集大小	5 925	3 080	6 429
文本平均长度	8.3	56.9	9.1
类别标签数量	150	77	64

（1）CLINC150[33]是目前意图分类任务中最广泛使用的数据集之一，用于评估自然语言理解系统在多领域、多种标签设置下的泛化能力。该数据集共包含150个意图标签，划分为10个领域，总计23 700条用户意图表达句。其中，每个意图包含150条训练样本和30条测试样本，文本平均长度为8.3个词。该数据集能够较好评估模型面对“细粒度意图区分”与“类间语义模糊”时的表现。

（2）Banking77[34]是金融领域主要涉及银行业务场景的意图分类数据集，也是金融领域智能客服系统研究的重要数据来源。包含77个银行业务相关的用户意图类别，共计13 083条查询句，每个意图大约包含100~200条实例。与CLINC150相比，Banking77数据集的文本表达更具任务导向性，语义集中度更高，但同样存在标签边界模糊、用户表述模糊等问题，本文选取其作为测试意图分类系统在具体的垂直场景下处理真实用户请求的能力。

（3）HWU64[35]是另一个常用的意图分类数据集，旨在模拟家庭助手与智能设备交互过程中的意图分类任务。该数据集包含64个常见意图标签，划分为多个智能场景（如天气、时间、设备控制、音频播放等），共计25 716条用户语句，平均文本长度为9.1个词。其数据构造方式是基于多轮对话的采样，具有自然语境下用户对智能设备进行询问请求的笼统、模糊语言特性，标签语义覆盖范围较广，非常适合验证本研究提出的联想语义建模机制的有效性。

4.2 基线模型与分析

为了多方面评估所提方法有效性，本文选取以下9种兼具代表性和先进性的基线模型进行对比：

（1）BERT[17]模型是直接采用预训练的BERT编码器对用户查询文本进行双向上下文编码，并在顶层添加一个全连接分类器实现对意图类别的判别。BERT在多种自然语言任务上表现优异，但在低资源或零样本场景下对标注数据依赖较大，是目前基础但重要的对比方法。

（2）Sentence-BERT[21]是在传统BERT基础上进行微调，引入双塔结构和对比损失，通过端到端的句子对训练，使得句子级表示在语义相似度计算中具有更高的效率和准确度。在本文的意图分类任务中，将用户输入和意图标签集合分别编码为语义向量输入给该模型进行意图分类任务训练，适用于快速检索和零样本推断场景。

（3）Parrot-T5[36]是在T5预训练模型的基础上引入了数据增强与重写任务，通过释义生成，获得与原始查询同义或相关的句子变体，并使用这些变体进一步微调T5分类器，以提高对多样化表达的鲁棒性。

（4）T-CVAE[37]是基于条件变分自编码器框架，在编码器端添加意图标签条件信息，学习查询文本潜在语义分布，并在解码器端生成增强样本或重构查询。该模型能够在数据稀缺的情况下利用生成式能力扩充训练数据。

（5）GPT-4 Zero-Shot[9]利用大语言模型GPT-4的零样本能力，我们将用户查询和意图标签集合通过最直接的意图分类提示词组合后输入GPT-4，直接引导模型以零样本形式直接生成输出最匹配的意图类别。该方法无需额外训练阶段，只依赖通用预训练知识，是用于对比测试本文模型设计框架有效性的一个重

要基准方法。

(6)IntentGPT[25]是一种基于GPT-4模型改进的提示学习方法,区别在于提示词的设计中融合了领域描述、示例意图和联想提示,增强了模型对具体任务的专用推理能力。也是通过引导模型以零样本形式直接生成输出最匹配的意图类别,在有限示例或无示例条件下均能提供稳定的意图判别结果。

(7)MetaTCN[26]是一种用于少样本文本分类的元学习框架。该框架利用三元组对比学习来同时优化双向知识嵌入,并采用动态变化率采样策略来优先处理困难样本。

(8)DLNR[27]是一种基于大语言模型的用于小样本对话意图分类的动态标签精化方法,其主要特点是其动态性和上下文感知能力,与静态方法相比,它能根据每个测试查询的具体上下文自适应地调整标签,从而更精准地捕捉特定语境下的语义关系,避免破坏原始意图间的语义关联。

(9)MSCLM[28]是一种多语义对比学习框架,着重于解决阅读理解中的表征不一致问题。它使用交叉注意力机制,并通过基于提示词的对比学习来对齐字面含义和隐喻含义,在包括GPT-3.5在内的多个基准测试中实现了有竞争力的性能。

需要指出的是,本研究的意图分类任务与基线方法的评估方式保持一致,均采用了三项具有代表性的分类性能指标:准确率(Accuracy)、召回率(Recall)与宏平均F1值(M-F1),以综合评估基线模型和L-ZCAM模型在不同数据集的性能表现。

4.3　实验设置

本研究在模型配置、训练参数与硬件环境三方面作了统一设定,均基于Python3.11.7、Pytorch2.4.0框架和Huggingface平台获取所有的数据集和模型,以保证实验的对比公平性。具体而言,本文主干大语言模型基于GPT-4接口,BERT编码器采用谷歌的BERT large model(uncased),隐藏层维度为$d=1\,024$;为实现轻量化微调,引入LoRA适配器[23],设置秩为8、注意力头数为16、滑动窗口大小取3,以捕获潜在的局部上下文特征;对于标签定义的提示词来说,中文提示词构造形式为:“你是意图分类任务领域的专家,接下来你需要基于以下[数据集的任务描述]和[标签集合],为每个标签生成相应的[一句摘要性定义描述],并以列表形式输出”;而对于联想意图的结构化提示词来说,需要按照数据集任务特点逐一构造提示词模板,翻译成中文,提示词构造形式大致为:“你是[数据集的任务描述]领域的专家,接下来你需要基于[询问输入文本]展开联想,分析输出潜在传达的[需求/实体/产品/业务……]意图,并为每个意图生成[一句摘要性描述],然后以列表形式输出”。

对于每个询问输入文本来说,我们针对L-ZCAM构建的联想层次数量(即聚类簇数)M进行了实验,以此探究超参M对于模型分类效果的影响。相应的实验结果如图4所示,可见随着M从1开始增加,L-ZCAM能够引入更多的文本外语境信息,分类效果也随之增加,但当M到达6之后,出现了模型效果反而降低的饱和迹象。经过调试,最终确定$M=5$为最合适的选择。此外,针对不同的用户对话场景,针对大语言模型提示工程的质量对于模型最终的意图分类效果来说有着重要影响。为此,从实验选用的三个数据集角度开展分析,设计并调试了细分指令,用于帮助大语言模型拆解语义空间,让它有针对性地、有方向性地进行零样本语境联想,生成相应数据集下有价值的语境信息,充当为近似于人类看待任何一件事物时已经掌握的知识经验储备。相应地,本文分析了CLINC150、Banking77、HWU64三个意图分类数据集的文本特点,用户所传达的意图可以抽象为若干认知维度,包括:需求(如:目标、命令)、实体(如:银行卡、灯光)、业务(如:存款、导航)、场景限制(如:时间、地点)、领域(如:银行业务、驾驶助手)等。

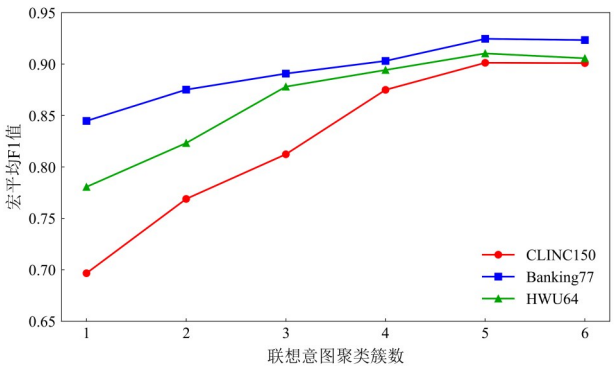

图4　L-ZCAM在不同联想意图聚类簇数下的性能表现

Figure 4　Performance of L-ZCAM under different numbers of associative intent clusters

具体来说,如式(4)所述,p_i^a为指定了需求、实体、业务等细分引导的指令,根据三个数据集的不同领域,相应的提示词指令也存在些许差异。经过调试,面向不同的用户对话场景,细分引导指令翻译成中文大体如下(联想层次数量/聚类簇数$M=5$):

(1)CLINC150数据集

“联想层次维度”:[“用户需求”,“目标实体”,“情境背景”,“领域行为”,“功能目标”],

“指令”:{

“用户需求”:“描述用户从根本上想要完成的事情”,

“目标实体”:“指出查询中涉及的对象或项目”,

“情境背景”:“说明该请求可能出现的现实世界情境”,

“领域行为”:“推断与该请求相关的典型操作或过程”,

“功能目标”:“描述用户可能试图实现的更深层目的”}

(2)Banking77数据集

“联想层次维度”:[“金融需求”,“银行相关实体”,“流程阶段”,“风险或约束”,“服务目标”],

“指令”:{

“金融需求”:“解释用户潜在的金融意图”,

“银行相关实体”:“指出涉及的银行卡、账户或银行产品”,

“流程阶段”:“推断该请求属于银行业务流程中的哪个步骤”,

“风险或约束”:“描述可能涉及的担忧、风险或紧迫性”,

“服务目标”:“描述用户希望获得的银行服务结果”}

(3)HWU64数据集

“联想层次维度”:[“用户指令”,“设备或应用实体”,“执行动作”,“环境语境”,“助手行为”],

“指令”:{

“用户指令”:“解释用户的主要命令或请求”,

“设备或应用实体”:“指出涉及的设备、应用或联系人”,

“执行动作”:“描述为满足该请求所需的系统动作”,

“环境语境”:“说明家庭、设备状态或其他上下文假设”,

“助手行为”:“说明智能助手在功能上应执行的任务”}

在模型的训练阶段,本文采用Adam优化器,初始学习率为2×10^{-5},权重衰减系数为0.01,训练批次大小为32,最大训练轮次为20,早停耐心训练轮次取5,并在表征层引入概率为0.3的Dropout机制以防止过拟合。硬件配置为Intel(R)Xeon(R)CPU E5-2650 v4@2.20 GHz,所有实验均在两块NVIDIA A100 80 GB GPU上以混合精度(FP16)进行,并结合DeepSpeed Zero-3策略进行显存优化,保证高效计算的同时能够最大化资源利用率。

4.4 主对比实验结果分析

在本节中,我们将系统分析L-ZCAM与现有主流意图分类方法在CLINC150、Banking77和HWU64三个标准数据集上的实验结果。表2展示了各模型在三个数据集上的性能对比,实验数据均基于三次独立运行取均值的形式呈现,其中(1~3)为传统的预训练模型,(4~8)则是基于生成内容或大语言模型的方法。

表2 不同模型方法在三种数据集上的性能对比 单位:%

Table 2 Performance comparison of different methods on three datasets unit: %

模型名称	CLINC150			Banking77			HWU64		
	Accuracy	Recall	M-F1	Accuracy	Recall	M-F1	Accuracy	Recall	M-F1
(1) BERT	68.54	65.21	62.18	71.23	68.15	66.75	74.85	71.92	70.47
(2) Sentence-BERT	78.34	75.67	73.27	82.15	79.82	78.04	80.27	77.43	75.91
(3) Parrot-T5	80.28	77.15	75.33	83.42	80.37	78.75	81.53	78.28	76.89
(4) T-CVAE	73.19	70.25	68.41	76.82	73.95	71.77	75.05	72.18	70.32
(5) GPT-4 Zero-Shot	78.72	75.83	73.85	81.25	78.95	77.33	80.18	77.24	75.17
(6) IntentGPT	85.33	82.45	80.82	87.46	84.88	83.15	86.04	83.27	81.89
(7) MetaTCN	90.15	87.24	85.37	93.28	90.45	88.92	91.89	88.95	87.43
(8) DLNR	90.27	87.51	85.85	93.77	91.02	89.50	92.22	89.74	88.12
(9) MSCLM	92.86	89.95	88.14	95.63	92.84	91.28	94.25	91.36	89.87
(10) L-ZCAM	94.56	91.92	90.12	96.73	94.25	92.45	95.24	92.67	91.03

整体而言,从表2的实验结果可以看出,L-ZCAM在三个数据集的所有评估指标(准确率、召回率、宏平均F1)上均取得最优结果,相较于当下最新方法MSCLM,L-ZCAM在CLINC150、Banking77、HWU64三个不同任务场景下公共数据集上的宏平均F1分数分别改善2.25%、1.28%和1.29%,说明所提出的基于大语言模型的零样本语境联想机制在三个意图分类任务中具有一致且较好的适用性与鲁棒性。从三种数据集任务特性来看,CLINC150是一个多场景、多领域的小样本任务集,覆盖150种用户意图,具有意图分布离散、类别粒度细的特点;Banking77则来自金融领域,常包含结构复杂、术语密集的文本信息;而HWU64相对结构稳定,意图层次较浅,偏向任务驱动型人机对话。L-ZCAM在这三种不同任务分布上的稳

定领先,表明面对跨领域迁移、未知标签干扰和表达多样性等现实问题时依旧能够保持稳定而优异的意图理解与分类能力。值得注意的是,与基础的GPT-4 Zero-Shot零样本分类模型相比,L-ZCAM在引入基于GPT-4的零样本语境联想生成机制并结合对比学习约束后,其整体的监督学习框架在各项指标上均取得了近20个百分点的显著提升,有力体现了本论文模型结构和设计思想的合理性。

接着,结合表3不同模型方法的能力对比角度来看,表2中9个基线模型和本文所提的L-ZCAM模型在三项评价指标上呈现出预期中的性能排序,能够较好地揭示其各自建模策略的优劣。需要指出的是,表3中的"语境扩展能力"代表了模型的"不依赖于数据集样本的文本外语境扩展能力"。最基础的BERT模型由于仅依赖上下文词向量进行文本建模,缺乏额外的语境扩展能力与标签语义理解机制,因此在三个数据集上均表现最差。Sentence-BERT和Parrot-T5通过引入句子级表示和数据增强机制,在语义匹配与表达多样性方面有所改进,但面对跨领域标签歧义和用户模糊表达仍存在显著局限。对于早期的生成式模型T-CVAE来说,其虽采用了数据生成的方法,具有一定的文本外语境信息扩展能力,但是其相较于大语言模型来说缺乏提示词引导,生成内容普遍较短且缺乏导向性,整体效果十分受限于模型结构以及训练数据规模,不稳定问题突出。作为当下大语言生成模型的龙头和标杆,GPT-4 Zero-Shot能够通过简单直白的自然语言提示直接"赋予数据集标签",即可对测试集实现零样本分类,在各数据集上F1值均在75%左右,体现出其良好的泛化能力、语义推理和分类能力基础。然而,GPT-4对提示设计高度敏感,在缺乏外部标签定义或联想指引等结构化提示词的情况下,则会导致其在垂直领域、细粒度意图区分方面存在性能瓶颈。另一方面,IntentGPT通过融合领域描述和示例提示,相较于GPT-4 Zero-Shot在意图分类的引导方面表现更优,语境信息的扩展能够带来更多有用的文本外信息,但其多轮提示设计带来的交互成本较高,且未充分结合多层次语境联想策略,是本文所提L-ZCAM模型的下位对比基线。进一步来说,MetaTCN、MSCLM等为近年来先进的端到端元学习与对比学习的有监督分类框架,其中MetaTCN侧重于样本选择与结构迁移优化,而MSCLM基于GPT3.5,并通过引入多语义对齐机制进一步提升理解能力,在三个数据集上均有着不错效果,反映出MSCLM基于GPT大语言模型架构的优势,即有监督分类框架辅以提示生成能力能对不同领域的意图判断有着较好的适应性。DLNR作为大语言模型加持的小样本学习方法,在未经大规模训练的条件下获得了与基于有监督对比学习的MetaTCN相匹敌的效果,这也充分论证了大语言模型的语境扩展能力在增强现有方法语义理解建模方面的强大作用。

表3　不同模型方法的能力对比

Table 3　Capability comparison of different methods

模型名称	大语言模型	对比学习	语境扩展能力
(1) BERT	×	×	×
(2) Sentence-BERT	×	√	×
(3) Parrot-T5	×	×	×
(4) T-CVAE	×	×	√
(5) GPT-4 Zero-Shot	√	×	√
(6) IntentGPT	√	×	√
(7) MetaTCN	×	√	×
(8) DLNR	√	×	√
(9) MSCLM	√	√	√
(10) L-ZCAM	√	√	√

然后,从是否结合大语言模型这一维度来看,实验结果进一步验证了大语言模型在意图分类任务中的潜力与边界。结合LLM进行模型设计的几种方法(8~10),相较于基于传统结构的方法(1~4)普遍具有更高的性能水平,说明大模型具备更丰富的世界知识、语言先验和推理能力,能够有效对抗用户查询短文本中常见的上下文缺失和表达歧义问题。但与此同时,这些基于LLM的方法效果仍受提示设计策略、标签引导方式以及语义控制机制的制约。特别地,本文设计的L-ZCAM将LLM作为零样本语境联想语义生成器,进一步引入显式标签定义监督和结构化对比学习建模,成功实现了从"语言生成"到"分类判断"的桥接与优化,充分挖掘大模型的语义先验潜力并转化为有监督分类任务中的零样本语境生成能力。这种策略不仅在性能上超越了仅靠提示工程的零样本分类模型,也优于传统的有监督学习框架,为未来LLM驱动下的意图分类提供了一个具有实证意义的结构化路径。

综上所述,L-ZCAM不仅相较于当前各类主流的分类模型取得了最优结果,其基于的"LLM零样本语境联想+有监督训练"特性更展示出极强的跨任务适应性、泛化能力与模型可扩展性,验证了语境联想机制、标签定义引导、对比学习约束对意图分类任务的重要价值。

4.5　消融实验结果分析

本节对所提出的L-ZCAM模型进行了系统的消融实验,以进一步验证模型中各关键模块对整体性能的贡献。表4展示了七种模型变体与完整L-ZCAM模型在CLINC150、Banking77以及HWU64三个标准数据集上的分类性能对比,评价指标涵盖准确率(Accuracy)、召回率(Recall)和宏平均F1(M-F1)。

表4 L-ZCAM在三种数据集上的消融实验结果 单位:%

Table 4 Ablation study results of L-ZCAM on three datasets unit: %

消融配置	CLINC150			Banking77			HWU64		
	Accuracy	Recall	M-F1	Accuracy	Recall	M-F1	Accuracy	Recall	M-F1
完整L-ZCAM模型	94.56	91.92	90.12	96.73	94.25	92.45	95.24	92.67	91.03
w/o LoRA微调	91.34	88.65	86.89	93.56	91.02	89.23	92.15	89.48	88.12
w/o $\mathcal{L}_2$	90.27	87.58	85.62	92.45	89.86	88.11	91.39	88.72	87.04
w/o $\mathcal{L}_3$	89.05	86.33	84.33	91.14	88.42	86.85	90.02	87.35	85.67
w/o $\mathcal{L}_2+\mathcal{L}_3$	86.88	84.12	82.15	88.95	86.28	84.33	87.64	85.01	83.02
w/o 残差特征融合	88.15	85.42	83.48	90.28	87.63	85.72	89.05	86.38	84.56
w/o 交叉注意力	87.42	84.68	82.75	89.63	86.92	85.08	88.27	85.65	83.78
w/o 所有模块(=BERT)	68.54	65.21	62.18	71.23	68.15	66.75	74.85	71.92	70.47

从整体效果来看,完整的L-ZCAM模型在三个数据集上均优于所有消融变体。这一结果清晰表明,L-ZCAM所提出的LoRA微调模块、残差特征融合结构、双路交叉注意力机制以及引入的语义对齐损失函数$\mathcal{L}_2$和$\mathcal{L}_3$对于意图分类性能的提升都具有不可替代的作用。尤其值得注意的是,无论在多场景(CLINC150)、高专业术语密集度(Banking77)还是偏任务驱动型(HWU64)场景下,完整模型均展现出较好的鲁棒性与泛化能力,说明所提出的设计思路具备广泛适应性。

进一步地,对比七种模型变体的特点,不难发现每个模块的移除都会在不同程度上导致性能下降,且表现差异与各模块的功能高度相关。其中,移除LoRA微调模块(w/o LoRA微调)后,三项指标在所有数据集上均出现约3个百分点的下降,说明LoRA轻量适配机制虽然只对模型微调部分参数,但在高效调优大语言模型表示能力方面效果方面不可或缺。移除语义对齐损失函数(w/o $\mathcal{L}_2$ 与 w/o $\mathcal{L}_3$)后性能下降更加明显,特别是当同时去除两者(w/o $\mathcal{L}_2+\mathcal{L}_3$)时,接近IntentGPT等基础大语言模型水平。这一结果明确表明,通过KL散度对联想语义与标签定义之间的对齐约束是推动语义显著收敛和提升分类一致性的关键机制。同样值得关注的是,去除残差连接和去除交叉注意力结构后,性能进一步下降。这表明,交叉注意力机制在建模查询文本与联想意图、标签定义之间的语义交互关系方面至关重要,而残差特征融合机制则在保证信息传递稳定性、减缓梯度消失、提升表达层次性方面发挥了积极作用。这两者共同构成了L-ZCAM深层语义建模的框架,能够有效强化文本、意图和标签语义三者之间的匹配关系与联动能力。

需要指出的是,模型设计期间并未在$\boldsymbol{R}_T$(文本)和$\boldsymbol{R}_D$(定义)之间添加直接的损失函数约束,这源于一个核心出发点:“意图识别的关键在于挖掘字面关键词特征外隐式的语境联想信息,而非传统分类系统文本字面内容与标签定义的直接强对齐”。因此,本文在损失部分中围绕$\boldsymbol{R}_A$(联想)为中心,施加了两方面结构性约束(即$\mathcal{L}_2$与$\mathcal{L}_3$)。具体原因还包括以下两方面:首先,在人机交互过程中,用户所提供的文本本身具有很高的信息熵和不确定性,直接添加$\boldsymbol{R}_T$与$\boldsymbol{R}_D$的对齐约束会强化字面特征和分类标签的对齐,弱化大语言模型生成的文本外联想内容$\boldsymbol{R}_A$的作用,使模型偏向论文引言中图1所述的例子——“特征词浅层匹配”,而非“深层意图推断”。这一点与我们构建“语境联想中间层”的初衷相悖。其次,在L-ZCAM中,模型通过$\boldsymbol{R}_A$连接文本侧与标签侧,使其在语境抽象层面建立起以$\boldsymbol{R}_A$为中心的语义坐标系。因此,我们选择在$\boldsymbol{R}_T$与$\boldsymbol{R}_A$、$\boldsymbol{R}_A$与$\boldsymbol{R}_D$之间施加两个独立的KL约束,使由LLM生成的联想空间不仅可以与字面空间相结合($\mathcal{L}_2$),还可以直接贴近标签定义空间($\mathcal{L}_3$),从而保证隐喻式的联想表征也具备分类能力。

结合各模块设计初衷再度回顾实验结果,可以更好地理解L-ZCAM在各消融实验中的表现逻辑。首先,LoRA模块作为大语言模型调参的低秩适配机制,是实现轻量微调与参数高效共享的关键设计,使得我们可以在不大规模更新模型全部参数的前提下获得上下文感知能力的显著提升。其次,$\mathcal{L}_2$通过约束联想意图语义与询问输入文本语义之间的语义匹配,$\mathcal{L}_3$则作为联想意图与标签定义间的KL散度对齐损失,旨在从语义分布角度促进两个外部生成文本之间的统一语义收敛,在形成自洽的标签指导机制中发挥桥梁作用。残差特征融合结构是吸收残差网络思想引入的一种优化机制,确保信息流在多个编码层间流动平稳,有助于避免特征退化现象。而交叉注意力机制的查询向量互换结构模拟了“人类先感知、然后联想、进而推理”的过程,通过两两交叉对齐加强三类输入之间的语义交互,有效提升语义表达完整性和判别力。

综上所述,本节消融实验从多个维度全面验证了

L-ZCAM模型结构设计的合理性与模块贡献的独立性。每一项结构性创新都对最终性能提升具有不可替代的价值,尤其在大语言模型参与构建的条件下,如何有效融合联想生成内容、标签语义信息与用户输入文本成为影响分类性能的关键。L-ZCAM通过引入结构化的提示词输入与多路语义融合结构,构建起一种类人式的意图理解路径,其消融实验结果为未来在零样本文本理解与人机语义协同任务中的模型设计提供了强有力的实证支撑与路径参考。

4.6　不同大模型配置下的实验结果分析

为了评估大语言模型质量对于所提L-ZCAM方法效果的影响,本文结合三个额外的开源大语言模型Llama3-8b、Qwen2.5-7b、Qwen2.5-1.5b开展了对比实验,结果如表5所示。由此可见,L-ZCAM在不同大语言模型加持下呈现出与语言模型规模、推理能力以及语义联想深度密切相关的显著梯度差异。总体而言,使用GPT-4作为语义生成核心时,L-ZCAM在三个数据集上均取得了最好的效果。这一结果充分验证了GPT-4在联想语义生成、任务理解和高阶抽象能力上的优势,使其在零样本语境中能够更准确地生成高质量的联想意图集合与标签定义,从而显著提升了L-ZCAM的语义对齐能力。究其原因,L-ZCAM的性能高度依赖于联想意图生成模块与标签定义生成模块的语义质量,而这类生成任务本质上要求模型具备较强的语义组合能力、跨领域迁移能力以及隐含任务目标推断水平,因此GPT-4在深层语义理解与知识联想方面的优势直接转化为显著的下游意图分类增益。换言之,L-ZCAM在GPT-4加持下能够构建更精确、更具判别力的语义关联特征,使得三个特征空间(原始文本、联想意图、标签定义)之间的对齐更加紧密,最终实现最优的分类效果。

表5　L-ZCAM使用不同大语言模型配置的实验结果　　单位:%

Table 5　Experimental results of L-ZCAM with different large language model configurations　　unit: %

模型名称	CLINC150			Banking77			HWU64		
	Accuracy	Recall	M-F1	Accuracy	Recall	M-F1	Accuracy	Recall	M-F1
L-ZCAM (w/ GPT-4)	94.56	91.92	90.12	96.73	94.25	92.45	95.24	92.67	91.03
w/ Llama3-8b	92.31	88.40	87.55	95.12	91.03	90.12	92.87	89.56	88.42
w/ Qwen2.5-7b	90.84	87.92	85.63	93.46	90.22	88.04	92.51	88.97	86.11
w/ Qwen2.5-1.5b	88.75	85.03	82.47	91.20	87.58	85.01	90.84	87.12	84.37

5　结束语

在本文中,我们提出了一种基于大语言模型的零样本语境联想模型(L-ZCAM),旨在提升有监督意图分类任务中对模糊、短文本的语义理解和分类精度。通过设计联想意图生成和标签定义集合生成策略,L-ZCAM能够有效地对用户输入的查询文本进行多层次的语义扩展和特征增广,既能够从文本外部的联想语境中汲取信息,又实现了“文本—联想—标签”三种信息空间有效的语义对齐,从而实现更加精准和鲁棒的意图分类。在实验部分,详细验证了所设计模型在三个公开数据集上的分类效果,并与当前主流的多种基线模型进行了对比。实验结果表明,L-ZCAM通过结合大语言模型的生成能力与零样本语境联想机制,能够有效突破传统有监督方法在低资源场景下的性能局限,能够在各种复杂、多样的用户对话场景下进行高效的语境联想和意图分类。未来的工作将进一步探索该模型在多模态、跨语言及对话系统等更复杂任务中的应用。

参考文献

[1] Weld H, Huang Xiaoqi, Long Siqu, et al. A survey of joint intent detection and slot filling models in natural language understanding[J]. ACM Computing Surveys, 2023, 55(8): 1-38.

[2] Hrycyk L, Zarcone A, Hahn L. Not so fast, classifier-accuracy and entropy reduction in incremental intent classification[C]//Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI. Stroudsburg: ACL, 2021: 52-67.

[3] Wang Haifeng, Li Jiwei, Wu Hua, et al. Pre-trained language models and their applications[J]. Engineering, 2023, 25: 51-65.

[4] Ferré P, Fraga I, Hinojosa J A. The interplay between language and emotion: A narrative review[J]. Cognition and Emotion, 2025, 39(7): 1418-1445.

[5] Wang Jindong, Lan Cuiling, Liu Chang, et al. Generalizing to unseen domains: A survey on domain generalization[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(8): 8052-8072.

[6] Gardazi N M, Daud A, Malik M K, et al. BERT applications in natural language processing: A review[J]. Artificial Intelligence Review, 2025, 58(6): 166.

[7] Tompkins V, Montgomery D E, Dore R A, et al. Theory of

mind and text comprehension across the lifespan: A meta-analysis[J]. Developmental Psychology, 2025, 61(6): 1112-1125.

[8] Bzdok D, Thieme A, Levkovskyy O, et al. Data science opportunities of large language models for neuroscience and biomedicine[J]. Neuron, 2024, 112(5): 698-717.

[9] OpenAI, Achiam J, Adler S, et al. GPT-4 technical report[PP/OL]. V6. arXiv (2024-03-04) [2026-01-04]. https://doi. org/10.48550/arXiv.2303.08774.

[10] Anil R, Dai A M, Firat O, et al. PaLM 2 technical report[PP/OL]. V3.arXiv (2023-09-13)[2026-01-04]. https://doi.org/10.48550/arXiv.2305.10403.

[11] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models[PP/OL]. V1.arXiv (2023-02-27)[2026-01-04]. https://doi.org/10.48550/arXiv.2302.13971.

[12] Chang Yupeng, Wang Xu, Wang Jindong, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-45.

[13] 赵健程, 冯良骏, 岳嘉祺, 等. 从零样本学习理论模型到工业应用: 动机、演变与挑战[J]. 控制与决策, 2024, 39(9): 2833-2857.
Zhao Jiancheng, Feng Liangjun, Yue Jiaqi, et al. From zero-shot learning theoretical model to its industrial application: Motivation, evolution and challenges[J]. Control and Decision, 2024, 39(9): 2833-2857. (in Chinese)

[14] Shah C, White R, Andersen R, et al. Using large language models to generate, validate, and apply user intent taxonomies[J]. ACM Transactions on the Web, 2025, 19(3): 1-29.

[15] Zhang Jiong, Chang Weicheng, Yu H F, et al. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification[J]. Advances in Neural Information Processing Systems, 2021, 34: 7267-7280.

[16] Saha Roy R, Katare R, Ganguly N, et al. Discovering and understanding word level user intent in Web search queries[J]. Journal of Web Semantics, 2015, 30: 22-38.

[17] 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). Kerrville: Association for Computational Linguistics 2019: 4171-4186.

[18] 许婷, 肖桐, 张圣林, 等. 基于LLM的日志故障诊断[J]. 电子学报, 2025, 53(4): 1123-1141.
Xu Ting, Xiao Tong, Zhang Shenglin, et al. Log fault diagnosis based on large language models[J]. Acta Electronica Sinica, 2025, 53(4): 1123-1141. (in Chinese)

[19] Wang Xiang, Huang Tinglin, Wang Dingxian, et al. Learning intents behind interactions with knowledge graph for recommendation[C]//Proceedings of the Web Conference 2021. New York: ACM, 2021: 878-887.

[20] 吴天舒, 尹宏鹏, 赵丹丹, 等. 基于迁移学习的零样本故障诊断[J]. 电子学报, 2023, 51(9): 2572-2577.
Wu Tianshu, Yin Hongpeng, Zhao Dandan, et al. Zero sample fault diagnosis based on transfer learning[J]. Acta Electronica Sinica, 2023, 51(9): 2572-2577. (in Chinese)

[21] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2019: 3980-3990.

[22] Mohsenimofidi S, Prasad A S R, Zahid A, et al. Classifying user intent for effective prompt engineering: A case of a chatbot for startup teams[M]//Nguyen-Duc A, Abrahamsson P, Khomh F. Generative AI for Effective Software Development. Cham: Springer Nature Switzerland, 2024: 317-329.

[23] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[C/OL]//Proceedings of the International Conference on Learning Representations, 2022: 1-13. https://openreview.net/forum?id=nZeVKeeFYf9.

[24] Tian Y J, Han Y K, Chen X S, et al. Beyond answers: Transferring reasoning capabilities to smaller LLMs using multi-teacher knowledge distillation[C]//Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining. New York: ACM, 2025: 251-260.

[25] Rodriguez J A, Botzer N, Vazquez D, et al. IntentGPT: Few-shot intent discovery with large language models[C]//Proceedings of the ICLR 2024 Workshop on Large Language Model Agents, 2024. https://openreview.net/forum?id=IDuQtpSgGp.

[26] Dong Kaifang, Jiang Baoxing, Li Hongye, et al. Meta-learning triplet contrast network for few-shot text classification[J]. Knowledge-Based Systems, 2024, 303: 112440.

[27] Park G, Baek I, Kim B, et al. Dynamic label name refinement for few-shot dialogue intent classification[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2025:

41-52.

[28] Wu Mingmin, Hu Yuxue, Zhang Yongcheng, et al. Mitigating idiom inconsistency: A multi-semantic contrastive learning method for Chinese idiom reading comprehension[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17): 19243-19251.

[29] Wong B I, Lecompte M, Yang L X. Associative memory with value-directed learning in younger and older adults[J]. Aging, Neuropsychology, and Cognition, 2025, 32(6): 891-906.

[30] Connell L, Lynott D. What can language models tell us about human cognition?[J]. Current Directions in Psychological Science, 2024, 33(3): 181-189.

[31] Xu Miao, Li Yufeng, Zhou Zhihua. Robust multi-label learning with PRO loss[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(8): 1610-1624.

[32] 王进, 刘彬, 孙开伟, 等. 基于标签关联的多标签演化超网络[J]. 电子学报, 2018, 46(4): 1012-1018.
Wang Jin, Liu Bin, Sun Kaiwei, et al. Multi-label evolutionary hypernetwork based on label correlations[J]. Acta Electronica Sinica, 2018, 46(4): 1012-1018. (in Chinese)

[33] Larson S, Mahendran A, Peper J J, et al. An evaluation dataset for intent classification and out-of-scope prediction[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2019: 1311-1316.

[34] Coope S, Farghly T, Gerz D, et al. Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 107-121.

[35] Liu Xingkun, Eshghi A, Swietojanski P, et al. Benchmarking natural language understanding services for building conversational agents[M]//Marchi E, Siniscalchi S M, Cumani S, et al. Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems. Singapore: Springer Singapore, 2021: 165-183.

[36] Zhao Chao, Vijjini A, Chaturvedi S. PARROT: Zero-shot narrative reading comprehension via parallel reading[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: ACL, 2023: 13413-13424.

[37] Wang Tianming, Wan Xiaojun. T-CVAE: Transformer-based conditioned variational autoencoder for story completion[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2019: 5233-5239.

作者简介

陶汉卿　男,1995年3月出生于安徽省宿州市。现为中国矿业大学信息与控制工程学院助理研究员。主要研究方向为人工智能、数据挖掘和自然语言处理。
E-mail: hqtao@cumt.edu.cn

程玉虎　男,1973年8月出生于安徽省淮南市。现为中国矿业大学信息与控制工程学院教授、博士生导师。主要研究方向为强化学习、具身智能。
E-mail: chengyuhu@163.com

王雪松　女,1974年12月出生于安徽省泗县。现为中国矿业大学信息与控制工程学院教授、博士生导师。主要研究方向为机器学习、人工智能。中国电子学会会员编号:E190006839S。
E-mail: wangxuesongcumt@163.com

王　军　男,1981年1月出生于江苏省徐州市。现为中国矿业大学信息与控制工程学院教授、博士生导师。主要研究方向为智能机器人与无人系统、生物特征识别、机器视觉。中国电子学会会员编号:E190089908M。
E-mail: jrobot@126.com