文章编号:1671-6833(2024)01-0078-12

# 基于小样本学习的口语理解方法综述

刘 纳[1,2], 郑国风[1,2], 徐贞顺[1,2], 林令德[1,2], 李 晨[1,2], 杨 杰[1,2]

(1.北方民族大学 计算机科学与工程学院,宁夏 银川 750021;2.北方民族大学 图像图形智能处理国家民委重点实验室,宁夏 银川 750021)

**摘 要**:小样本口语理解是目前对话式人工智能亟待解决的问题之一。结合国内外最新研究现状,系统地梳理了口语理解任务的相关文献。简要介绍了在非小样本场景中口语理解任务建模的经典方法,包括无关联建模、隐式关联建模、显式关联建模以及基于预训练范式的建模方法;重点阐述了在小样本口语理解任务中为解决训练样本受限问题而提出的基于模型微调、基于数据增强和基于度量学习 3 类方法,介绍了如 ULMFiT、原型网络和归纳网络等代表性模型。在此基础上对不同模型的语义理解能力、可解释性、泛化能力等性能进行分析对比。最后对口语理解任务面临的挑战和未来发展方向进行讨论,指出零样本口语理解、中文口语理解、开放域口语理解以及跨语言口语理解等研究内容是该领域的研究难点。

**关键词**:口语理解; 小样本学习; 模型微调; 数据增强; 度量学习

**中图分类号**:TP391.1 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2024.01.012

近年来,对话式人工智能(dialogue artificial intelligence,DAI)在工业、医疗、金融和教育等领域受到广泛的关注。DAI 是一种能够进行自然语言对话的人工智能技术,通过将自然语言处理(natural language processing,NLP)、语音识别(automatic speech recognition,ASR)[1]、语义理解和对话理解等技术应用到智能语音对话系统中,以实现实时有效的人机交互。根据 DAI 的应用场景,将其划分为面向任务的对话系统(task-oriented dialogue,TOD)和开放域对话系统(open-domain dialogue,ODD)两大类。其中,TOD 主要解决针对某一具体领域的问题。例如,医疗行业部署智能对话系统完成病情分析、药品信息查询和提供诊疗方案等任务;教育领域利用智能对话系统实现教学体验提升、定制学习方案和获取学习资源等业务;金融领域则利用智能对话系统提供账户余额查询、定制理财方案等服务。ODD 需要实现与人类建立情感联系,进行共情对话。与 TOD 相比,ODD 的对话主题更为开放、覆盖范围更广、实现难度更大,是对话式人工智能亟待发展的研究方向之一。

2022 年 11 月,OpenAI 公司发布了全新的对话式通用人工智能工具即 ChatGPT,受到了全球各界的广泛关注。ChatGPT 产品的落地标志着大规模预训练语言模型(pre-train language model,PLM)已经具备了通用人工智能的特征。在 ChatGPT 产品问世之后,OpenAI 公司于 2023 年 3 月发布了最新的语言模型 GPT-4,其性能与 ChatGPT 最初使用的 GPT-3.5 模型相比有了巨大的提升。在口语理解方面,模型的理解能力、回答的可靠性有了显著提高。

中国类似于 ChatGPT 的研究也正在进行,例如百度公司推出了基于文心大模型的产品文心一言;复旦大学发布了中国第一个对话式大型语言模型 MOSS;在教育领域网易公司将类 ChatGPT 技术进行落地研发等。目前,中国在通用人工智能领域的发展与外国相比还有很大的差距,但发展速度快,与国际领先水平的差距会随着对大规模预训练语言模型的持续研究而逐渐缩小。

目前针对口语理解任务的研究综述较多,2020 年,Louvan 等[2]根据神经网络结构特征对口语理解任务的方法进行归纳。2022 年,Weld 等[3]针对如何提高联合模型的能力、如何捕获深层次语义和如何提高模型的泛化性 3 大问题,对前人的工作进行

**收稿日期**:2023-08-20;**修订日期**:2023-09-28

**基金项目**:宁夏自然科学基金资助项目(2021AAC03224,2021AAC03217)

**作者简介**:刘纳(1986—),女,宁夏银川人,北方民族大学讲师,博士,主要从事数据挖掘与自然语言处理技术研究,E-mail:liuna@nun.edu.cn。

**引用本文**:刘纳,郑国风,徐贞顺,等. 基于小样本学习的口语理解方法综述[J]. 郑州大学学报(工学版),2024,45(1):78-89.(LIU N, ZHENG G F, XU Z S, et al. A survey of spoken language understanding based on few-shot learning [J]. Journal of Zhengzhou University (Engineering Science),2024,45(1):78-89.)

总结。但以上大多数研究都采用非小样本学习的方法,对研究者来说,获取大量有标注的训练样本代价非常昂贵,并且对于新出现的意图领域,带标注的样本较少,获取十分困难。与之前的工作相比,本文主要对在小样本场景中口语理解任务的建模方式进行介绍,具有较强的针对性。

本文首先简要介绍了在非小样本场景中,口语理解任务建模的经典方法;其次,重点阐述了在小样本口语理解任务中为解决训练样本受限问题而提出的基于模型微调、基于数据增强和基于度量学习 3 类最新研究方法,并对不同方法的优缺点进行全面的比较与总结归纳;最后,对小样本口语理解领域存在的问题与挑战进行分析。

## 1　相关工作

口语理解(spoken language understanding, SLU)是对话式人工智能系统的核心任务之一。它的目标任务是提取用户输入的话语中所包含的意图,即用户的行为,并给予一定的反馈。2011 年,Tur 等[4]将口语理解任务划分为意图分类和槽位填充两个子任务。如表 1 所示,在槽位填充任务中采用的是 BIO 标注方案,通过意图分类识别用户的具体行为。

**表 1　口语理解任务举例**

**Table 1　Examples of spoken language understanding tasks**

| 任务 | 样本 | | | | |
|---|---|---|---|---|---|
| | Book | a | flight | to | Beijing |
| 槽位填充 | B | I | I | O | B |
| 意图分类 | BookTicket | | | | |

根据两个子任务之间的关联程度将非小样本场景下的口语理解相关研究划分为 4 类:①无关联建模,意图分类与槽位填充任务分别单独建模;②隐式关联建模,意图分类与槽位填充联合建模,获取两个子任务之间的全部共享信息;③显式关联建模,意图分类与槽位填充联合建模,获取两个子任务之间有用的共享信息;④基于预训练范式建模,以上下文感知为核心,捕获更深层次的语义信息。

### 1.1　无关联建模

无关联的建模方式将口语理解任务划分为意图分类和槽位填充两个子任务单独建模,模块化设计让每个模型结构简单、灵活,并且可以在不修改其他模块的情况下对特定的任务进行调整。

2013 年,Bhargava 等[5]对口语理解任务单独建模进行了早期尝试。利用支持向量机(support vector machine, SVM)对意图分类任务建模,利用条件随机场(conditional random field, CRF)对槽位填充任务建模。同时结合上下文信息,将前一个话语中的知识合并到当前话语中,显著提高了意图分类与槽位填充任务的性能,这是口语理解任务无关联建模的开端。

随着深度学习的发展,循环神经网络(recurrent neural networks, RNN)表现出强大的语言建模能力。2015 年,Mesnil 等[6]采用 RNN 对槽位填充任务进行了深入研究,比较了 RNN 的几种变体,其中包括 Elman-type 网络和 Jordan-type 网络。在 ATIS 数据集上,两种网络结构的性能都优于 CRF 模型。2017 年,Lin 等[7]认为基于 RNN 的递归模型在所有的时间步中携带样本的语义信息非常困难,并且会造成灾难性遗忘的问题,因此对传统的句子编码方式进行改进,设计双向 LSTM 结构,使用自注意力机制替换传统的最大池化或平均池化,从而有效减少了 RNN 的长期记忆负担。

卷积神经网络(convolutional neural network, CNN)最初应用在图像领域中,后来研究者将 CNN 应用在语义融合、句子建模等 NLP 任务中,同样取得了非常出色的效果。2014 年,Kim[8]在 word2Vec 基础上添加了卷积神经网络结构,使用词向量嵌入与 CNN 相结合的方式进行文本分类任务。CNN 利用不同大小的卷积核来提取句子中的关键信息,更好地建立局部语义相关性。但其存在的缺陷是难以提取对于距离大于卷积核窗口长度的特征,同时使用最大池化仅保留提取特征向量的最大值,导致部分重要的位置编码信息丢失。针对上述 CNN 的缺陷,2018 年,Zhao 等[9]开启了使用动态路由的胶囊网络进行文本分类任务的早期探索。胶囊网络利用神经元向量替代传统神经网络的单个神经元节点,显著改善了 CNN 空间不敏感的问题。利用动态路由算法调整子胶囊与父胶囊之间的权重,解决了使用最大池化算法丢失位置编码信息的问题。

无关联的建模方式存在的缺陷是需要对每个任务进行单独建模,模型结构整体较为庞大。各任务的模型之间没有数据或功能共享,易产生数据碎片。在实际的应用场景中,某些意图和槽位信息会在多个领域之间共享,无关联的建模方式无法利用两个任务之间的共享知识,导致用户在与系统交互过程中达不到满意的效果。为解决上述问题,后续工作提出了联合建模的方法。

### 1.2　隐式关联建模

联合建模思想的提出,极大地促进了口语理解领域的研究。但在早期的工作中,大多数采用隐式

联合建模的方式。仅通过共享编码器(shared encoder)捕获意图分类和槽位填充两个子任务之间的共享特征,之间没有进行任何的显式交互。

2016 年,Zhang 等[10]首次提出将意图分类与槽位填充任务进行联合建模,并首次将 RNN 结构引入到意图分类任务中。由于 RNN 对于捕获长期依赖关系十分困难,同时会带来梯度消失和梯度爆炸等问题,因此选择基于 RNNs 改进的门控循环神经网络 GRU[11]作为模型的基础架构。该联合模型的缺陷在于需要等待输入序列全部输入到模型之后才能开始后续的意图分类任务,实时性差。在实际的 SLU 应用中,用户对系统的实时性要求通常较高。为解决上述问题,Liu 等[12]提出基于 LSTM 的联合 SLU 实时模型。由于 LSTM 具有较强的捕获词序列中长期依赖关系的能力,因此使用 LSTM 作为基本的 RNN 单元。通过对整个序列上的 RNN 单元输出取平均值作为样本的表示向量,利用最后一个 RNN 单元输出预测的意图类别。对当前时间步以及之前时间步的隐藏状态建模槽位标签之间的依赖关系,每个时间步以单个词语作为输入,输出对应的槽位标签。Liu 等[13]借鉴注意力机制在机器翻译领域的成功经验,首次提出将基于注意力机制的循环神经网络模型应用在联合意图分类和槽位填充任务中。与机器翻译不同的是,在槽位填充任务中,输入的文本与输出的标签之间具有一一对应的关系,因此采用 Seq2Seq 结构,如图 1 所示。编码层使用双向 LSTM,可更好地捕获长期依赖关系。解码层使用 LSTM 并添加注意力机制预测槽位标签,在最后的隐藏层上通过前馈神经网络输出意图类别。

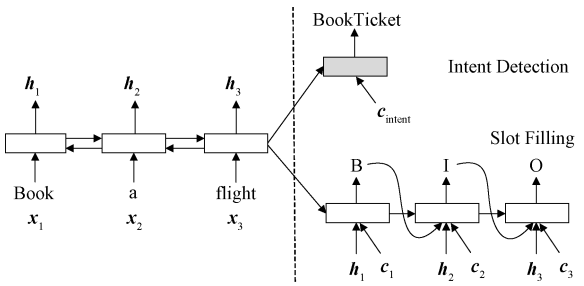


**图 1 Seq2Seq 建模口语理解任务结构图**

**Figure 1 Seq2Seq modeling spoken language understanding task structure diagram**

上述隐式联合建模的方式在一定程度上利用了意图分类和槽位填充两个任务之间的共享信息,极大地提高了口语理解的准确性。但缺陷在于隐式联合建模缺乏噪声过滤机制,两个子任务的噪声会在联合模型中进行传播,导致模型性能受限。为解决上述问题,后续工作提出了显式关联建模的方法。

### 1.3 显式关联建模

为解决隐式关联建模中的噪声传播问题,一些工作利用显式联合建模的方法,通过添加类似于门控机制的方式,选择性地获取意图分类和槽位填充任务之间的共享信息。

2018 年,Goo 等[14]首次提出使用显式建模的方式在意图分类与槽位填充两个任务之间建立联系。Goo 等[14]认为槽位信息通常高度依赖于意图信息,因此提出一种槽位门控机制(SGM-SLU),对意图与槽位注意力向量之间的显式关系进行建模。具体而言,在槽位门控模型中引入附加门,结构如图 2 所示。首先利用权重矩阵 $\boldsymbol{w}$ 将槽位向量 $\boldsymbol{C}_i^{\mathrm{S}}$ 与意图向量 $\boldsymbol{C}^{\mathrm{i}}$ 维度扩充一致,进行相加操作。接着经过槽位门控,在最后一个时间步中进行求和,得到 $\boldsymbol{g}$ 向量表示联合向量 $\boldsymbol{C}_i^{\mathrm{S}}$ 与 $\boldsymbol{C}^{\mathrm{i}}$ 的加权特征,其中 $\boldsymbol{g}$ 表示槽位与意图之间的关联程度。

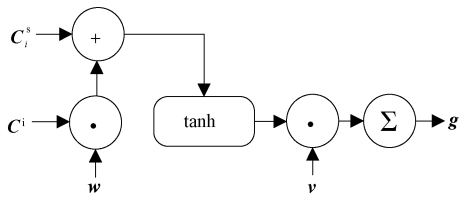


**图 2 SGM-SLU 结构图**

**Figure 2 Structural diagram of the SGM-SLU**

2019 年,Qin 等[15]认为 Goo 等[14]提出的仅依靠门控机制获取意图信息是有风险的,并且意图信息引导槽位填充任务具体过程的可解释性很差。因此,提出以堆栈作为数据结构的传播模型,将意图信息直接作为槽位填充任务的输入,提高了模型的可解释性。Chen 等[16]提出了一种具有条件随机场和先验掩码的多头自注意力联合模型。该模型使用多头局部自注意力机制来提取共享特征,使用掩码门控机制来建立意图分类和槽位填充两项任务输出的相关性,并使用 CRF 约束槽位填充任务的输出,充分利用了两个任务之间的语义关系。

以上的工作主要通过在意图分类与槽位填充任务之间建立单向交互,共享信息从意图流向槽位,无法充分利用它们之间的双向交互知识。Wang 等[17]设计了一种基于双模型的 RNN 语义框架解析网络结构,通过两个双向的 LSTM(BiLSTM)结合意图分类与槽位填充两个任务之间的双向交互知识,为每个样本同时生成意图和语义标签,显著提高了模型的性能。

基于以上工作可以发现传统基于 RNN 的方法

只能处理一定的短期依赖,无法处理长期依赖问题。后来基于 LSTM 和 BiLSTM 的模型结构在一定程度上突破了序列模型的局限性,但固有的顺序性限制了样本的并行化训练。显式联合建模的方式进一步利用了两个任务之间的共享知识,但模型无法捕获更深层次的语义信息。预训练模型的发展给口语理解任务带来了新的研究思路。

### 1.4　基于预训练范式建模

自然语言处理领域中的预训练研究思路最早可以追溯到 word2Vec 模型的提出。预训练的核心在于使用大量的训练数据,从中提取共性特征,帮助 NLP 下游任务简化其训练过程。早期的预训练模型专注于词向量编码,模型的特点是上下文无关,模型只知"上文"不知"下文",缺乏双向交互能力,代表性的工作包括 word2Vec、GloVe 等。近几年的预训练模型以上下文感知为核心,共享知识在上下文之间进行双向流动,代表性的工作包括 ELMo、BERT、GPT 等。

2019 年,Chen 等[18]首次将预训练模型应用到口语理解任务中,使用 BERT 预训练模型对意图分类和槽位填充任务进行联合建模,提出了 JointBERT 模型。模型结构如图 3 所示,BERT 预训练模型的下游任务之一是文本分类,因此很容易就能扩展到意图分类任务中。将[CLS]标签的输出替换成意图分类器,为后续的标签添加序列标签器,输出槽位最佳的标签匹配序列。槽位标签的预测取决于上下文单词的预测,由于结构化预测模型可以提高槽位填充的性能,在 JointBERT 模型的基础上添加 CRF 来对槽位标签之间的依赖关系进行约束建模。JointBERT 模型充分利用两个子任务之间的联系,捕获两个任务之间的共享知识。

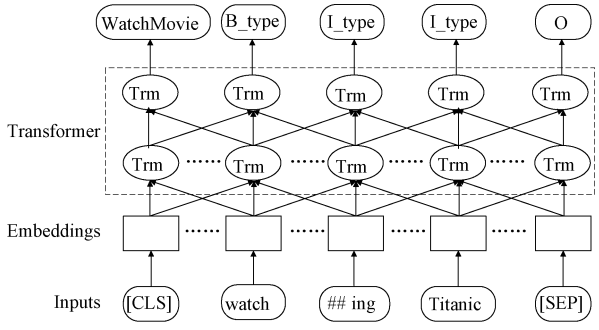


**图 3　JointBERT 模型结构图**

**Figure 3　JointBERT model architecture**

2020 年,Qin 等[19]认为仅识别对话中的显式意图并不能捕获用户的全部语义,对话中的隐式意图是更为重要的语义获取来源,因此提出一种协同交互式图注意力网络(Co-GAT)来联合对话显式意图分类和隐式意图分类这两项任务。模型的核心是设计一个协同的图交互层,可以同时获取上下文信息和交互信息。这是首次将上下文信息和交互信息结合进行联合显隐式意图识别的研究。

以上基于预训练语言模型的建模方式极大地促进了口语理解领域的发展。但通过对这些预训练模型性能的评估可以发现,目前基于预训练的方法并不能从根本上解决现有模型可解释性弱、泛化能力差、推理能力不足等问题,在深层次语义获取与理解方面还远远落后于人类的认知水平。同时,如何对大规模预训练语言模型进行压缩、降低参数量是一个亟待解决的问题。

## 2　小样本学习

早在 2006 年,Li 等[20]首次提出了小样本学习的概念。小样本学习致力于解决数据受限的深度学习问题,通过对少量样本甚至一个样本的训练使模型性能达到甚至超越大数据深度学习的效果。在生活中,有很多场景都属于小样本学习的范畴,例如儿童仅通过几张绘图卡片就能认识海洋生物,依靠少量的样本完成自主推理的过程。受到人类快速学习能力的启发,早期的研究人员将小样本学习方法应用在图像领域,解决训练样本数量受限的问题。2015 年,Koch 等[21]设计孪生神经网络解决了 one-shot 图像分类问题。Zhang 等[22]在关系网络的基础上,通过数据增强的方法解决了小样本图像分类问题。在自然语言处理领域,小样本学习发展较为缓慢,原因是图像特征相比于文本特征更为客观,在少量样本的情况下,提取文本特征更为困难。

近年来,随着预训练模型的发展,小样本学习在自然语言处理领域也有了一些突破。2018 年,Chen 等[23]使用对比学习框架解决小样本文本分类中的区分表示和过拟合问题。Jian 等[24]使用伪标签克服小样本学习固有的数据稀缺问题。以上方法在一定程度上缓解了由于数据过少无法支撑模型学习到足够的参数,在训练集上容易过拟合的问题。但大多数工作只专注于在已知的数据集上提高模型的学习上限,对于口语理解任务来说,注重的是模型对自然语言的理解与认知,而非学习浅层次的语义,这对模型的知识获取能力提出了更高的要求。

### 2.1　问题定义

在通常情况下,意图分类被看作是文本分类任务,将文本分类到指定的某个或者多个类别中。从数学定义上看,定义包含 $m$ 段文本的集合 $T=\{t_1,$

$t_2,\cdots,t_m\}$ 和包含 $n$ 个类别标签的集合 $C=\{c_1,c_2,\cdots,c_n\}$。模型最终产生由集合 $T$ 到集合 $C$ 的一对一或一对多映射关系。槽位填充被看作是序列标注任务,定义输入样本 $X=\{x_1,x_2,\cdots,x_n\}$,$x_i$ 表示样本中的某个字词,模型输出 $Y=\{y_1,y_2,\cdots,y_n\}$,$y_i$ 表示槽位标签。在小样本场景中,假设支持集 $S$ 包含 $N$ 种意图类别,每种意图类别由 $K$ 个样本组成,则将该任务称为 $N$-way $K$-shot 意图分类任务。

在近些年的研究中,基于小样本学习的口语理解方法主要分为 3 类:①基于模型微调的方法,将在大规模数据集上训练的模型迁移到目标任务中进行微调;②基于数据增强的方法,通过增强样本空间特征,提高模型的泛化能力;③基于度量学习的方法,利用度量函数计算样本之间的相似性。

## 2.2 基于模型微调的小样本口语理解

2015 年,Dai 等[25]首次提出了对语言模型进行微调的思想,模型需要先在大规模数据集上从 0 开始预训练,其次在小样本目标数据集上对全连接层或顶端神经网络结构的参数进行微调。该时期的微调模型经过海量数据的预训练才能表现出良好的性能,严重限制了模型的适应性。2018 年,Howard 等[26]提出了一种通用微调语言模型(universal language model fine-tuning,ULMFiT)。ULMFiT 模型训练主要由 3 个步骤组成:①在通用领域语言模型中进行预训练;②在目标任务语言模型中进行微调;③在目标任务分类器上进行微调。与其他模型的区别在于 ULMFiT 通过判别微调让模型的不同层学习不同的学习率。对于模型的同一层,随着迭代次数变化,使用倾斜三角学习率让参数进行自适应。判别微调与倾斜三角学习率机制让模型在小样本数据集上加快收敛速度,同时学习到更加符合目标任务的知识。

在 BERT 模型提出之前,传统的双向语言模型是将两个单向语言模型进行组合,而 BERT 模型是第一个基于微调的表示模型,在大型通用语料库中利用掩码语言模型(masked language model,MLM)和下一句预测任务(next sentence prediction,NSP)进行预训练。它使一系列 NLP 任务实现了当时最优的性能,表现出微调方法的巨大优势。2019 年,Sun 等[27]在 BERT 模型的基础上,研究如何通过微调 BERT 模型以解决长文本预处理、灾难性遗忘、低资源学习等问题。类似的工作还有 2020 年 Mohammadi 等[28]比较了微调不同层结构对 BERT 模型性能的影响,提出了 5 种不同的微调结构,如图 4 所示。

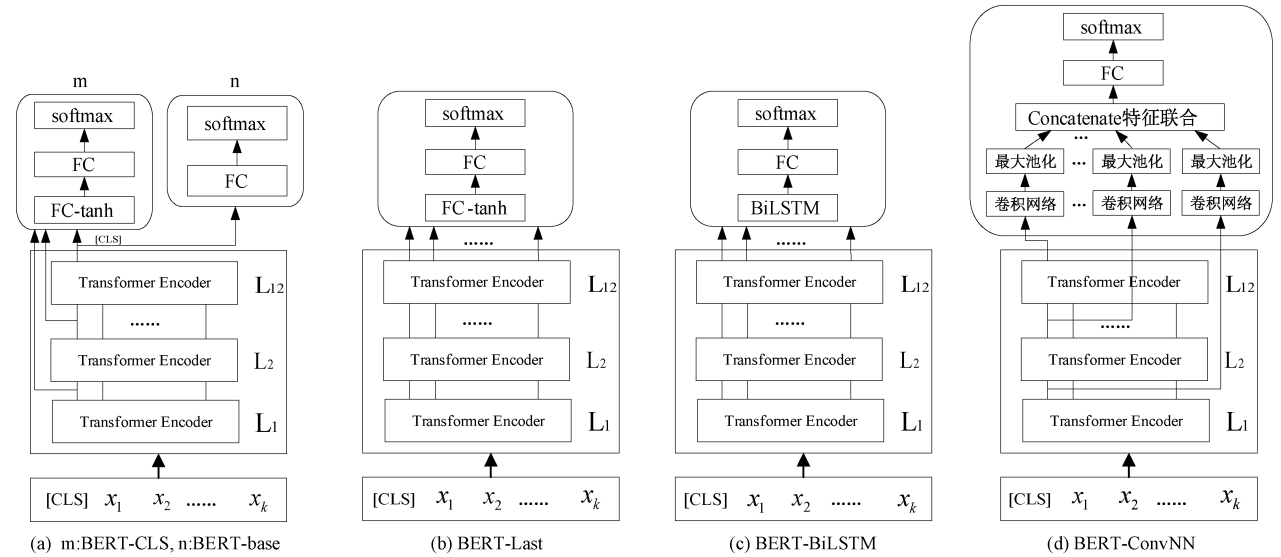


**图 4 5 种 BERT 模型微调结构图**

**Figure 4 Diagram of five fine-tuned BERT models**

如表 2 所示,在 30k-Intent(意图分类数据集,由 American Online 中 30 000 条用户检索意图文本组成)数据集上,Mohammadi 等[28]经过实验证明,BERT 模型通过微调添加 BiLSTM 与基础模型效果相似;在 BERT 模型之上添加全连接层作为分类器,可以得到最优的性能;卷积神经网络虽然具有较为复杂的网络结构,但该类结构不仅无法提高模型的精度,甚至会导致模型性能降低。

相比于 Mohammadi 等[28]的微调模型结构,2021 年 Zhang 等[29]用少量意图分类标注样本微调 BERT 模型,提出了一种新的微调模型 IntentBERT。该模

表 2　5 种微调模型结构对比

Table 2　Comparison of five fine-tuning models

| 模型名称 | 微调结构描述 | 微调目的 | 实验结果(*Accuracy*) |
|---|---|---|---|
| BERT-base | 添加全连接层 | 验证 BERT 模型的输出对分类任务效果的影响 | **0.64** |
| BERT-CLS | 所有隐藏层的第一个[CLS]作为全连接层的输入 | 将所有层的[CLS]作为分类器的输入,验证其对分类器性能的影响 | 0.60 |
| BERT-Last | Transformer 最后一层的全部输出作为全连接层的输入 | 验证[CLS]的有效性,使用最后一层的全部输出进一步提升分类模型的精度 | 0.62 |
| BERT-BiLSTM | 使用 BiLSTM 替换全连接层 | 验证 BiLSTM 的效果是否比全连接层更优 | **0.64** |
| BERT-ConvNN | 添加卷积神经网络 | 验证卷积神经网络能否提高模型的分类性能 | 0.43 |

型的优势在于目标领域的样本即使与预训练数据差异较大,也可以直接应用在目标领域上的小样本意图分类任务中,无须对目标数据进一步微调。但 IntentBERT 具有较强的各向异性(anisotropy),语义向量之间的余弦相似度较大,不同的语义难以分离。针对各向异性的问题,2022 年 Zhang 等[30]利用各向同性(isotropy)技术调整语义空间,通过调整目标函数的正则化项实现对模型的微调,提出了两种正则化项:①基于对比学习的正则化;②基于相关矩阵的正则化。实验证明,两种正则化相结合的微调方式在 BANKING77 和 HWU64 数据集上能够表现出更加出色的效果。

如表 3 所示,基于模型微调的方法思路简单,但在真实的应用场景中,预训练与微调之间的数据集和模型结构会产生显著偏差,导致微调的效果和预训练的效果会存在较大的差异性。同时,随着预训练语言模型的参数量呈现爆炸式增长,在下游任务上进行模型微调代价十分昂贵且耗时。为解决上述问题,后续工作提出了基于数据增强的方法和基于度量学习的方法。

表 3　基于模型微调的小样本口语理解模型对比

Table 3　Comparing few-shot spoken language understanding models with model fine-tuning

| 模型 | 基础模型 | 数据集 | 贡献 | 缺陷 |
|---|---|---|---|---|
| ULMFiT[26] | ULMFiT | TREC-6 | 第一个通用微调语言模型 | 易受到目标数据集与源数据集分布差异的影响 |
| 文献[27] | BERT | TREC | 减少了灾难性遗忘、低资源等问题对模型性能的影响 | 易导致在目标数据中过拟合 |
| 文献[28] | BERT_base | 30k-Intent | 比较了微调不同层结构对 BERT 模型性能的影响 | 对参数量巨大的预训练模型微调,代价昂贵 |
| IntentBERT[29] | BERT | BANKING77 | 使用小样本数据进行微调 | 具有较强的各向异性 |
| 文献[30] | BERT | BANKING77 | 解决了各向异性问题 | 正则化项增大了计算开销 |

### 2.3　基于数据增强的小样本口语理解

数据增强是通过增加样本的数量或空间特征,从而提高模型的泛化能力,缓解数据不足的问题。现阶段,NLP 领域的数据增强方法主要有:随机噪声注入、词汇替代、回译等。

2016 年,Kurata 等[31]首次将数据增强的思想引入到对话口语理解任务的模型中,利用编码器-解码器架构对训练样本中的数据进行重构。在数据增强的过程中,对编码器的输出隐藏层添加随机噪声来产生不同的样本,该方法的缺陷是增强产生的单个样本与其他样本之间没有建立关系。2018 年,Hou 等[32]针对该缺陷提出了一种新的数据驱动架构,对训练数据中相同语义框架的样本之间的关系进行建模。为了让生成的样本具有多样性,以 Seq2Seq 模型作为架构的核心,在样本表示中添加多样性等级队列(diversity rank),提升了生成样本的多样性并过滤相似的样本,显著提高了语言模型在标记数据稀缺领域的性能。

数据生成的方法在一定程度上避免了模型过拟合,生成的样本扩充了训练样本的数量,但缺陷在于模型会消耗额外的内存来生成噪声数据。2019 年,Kim 等[33]针对该缺陷提出了基于槽位添加噪声的方法,将数据转换成具有相同上下文、但不同槽位标签的短句来扩充数据。具体而言,对输入的训练数据进行噪声处理后,训练数据转变为包含噪声的嵌入向量,接着使用上下文作为神经网络的输入。模

型在每一步训练中使用不同的噪声数据,由于数据增强在相同的嵌入空间中执行,因此不需要花费额外的内存空间。

2019年,Zhao等[34]提出构造原子模板(atomic templates)进行数据增强。原子模板生成细粒度更好的语义样本,每一个模板由act-slot-value三元组组成。该方法的优势在于建立起act-slot-value三者之间的关系,而不是单独地对槽位或行为建模。在输入到句子生成器之前,用自然语言处理对话行为,以便生成器能够理解,提高了句子生成器的领域自适应能力。原子模板是在句子级上进行创建,减轻了人为创建模板的工作量。

为了提高口语理解模型的可变性和准确性,2021年,Peng等[35]提出基于预训练语言模型的数据增强方法,将在预训练阶段学习到的语法和语义融合到特定领域样本生成的过程中,该数据增强框架对生成的样本语义可控性更强。Qin等[36]基于预训练模型提出一种新的数据增强框架CoSDA-ML,用于生成多语种code-switching数据微调mBERT模型。该模型的主要思想是通过融合上下文信息来将源语言和多个目标语言的表示进行对齐。为了验证所提出的动态增强机制的有效性,与静态增强方法进行比较。模型的优势在于动态采样允许模型将更多的单词表示在多种语言中进行更紧密的对齐,同时对语言的依赖性较低,与mBERT模型相比,在各项NLP任务上的性能都有显著提高。

2022年,Sahu等[37]提出使用预训练语言模型生成意图样本,对任务进行数据增强。该方法的缺陷在于未考虑到生成样本的质量,模型可能会在低质量的生成样本上过拟合,同时生成的样本需要进行人工标记,成本较大。为解决上述问题,2023年,Lin等[38]引入Pointwise V-information(PVI)作为衡量过滤意图分类数据的指标,提出了基于PVI的上下文数据增强方法(in-context data augmentation,IC-DA),该方法首先在小部分训练数据上微调模型,接着在与已知意图相对应的样本上生成新的样本。经过实验证明,在BANKING数据集上,基于PVI的方法相比于未添加PVI过滤时意图分类准确率提高了4.45%。

对现有基于数据增强的方法进行分析和总结如表4所示。数据增强的方法通过增加样本的数量或空间特征,一定程度上提高了模型的泛化能力,但这些方法也会存在一些缺陷。例如:生成样本的质量会对模型产生影响,并且增强过程中可能会丢失一些关键信息。为了克服上述缺陷,一些工作转变研究思路,提出了度量学习的方法。

**表4 基于数据增强的小样本口语理解模型对比**

**Table 4 Comparing few-shot spoken language understanding models with data augmentation**

| 分类 | 模型 | 基础模型 | 数据集 | 贡献 | 缺陷 |
|---|---|---|---|---|---|
| 数据生成 | 文献[32] | Seq2Seq | ATIS、Stanford dialogue | 建立了样本之间的关联,提升了生成样本的多样性 | 生成样本的可解释性较弱 |
| | 文献[33] | BiLSTM | ATIS、SNIPS、MIT-Restaurant | 减少了额外内存空间的消耗 | 无法生成细粒度更好的样本 |
| 构造模板 | 文献[34] | BiLSTM | DSTC 2&3 | 原子模板生成细粒度更好的样本 | 模型对模板依赖性较大 |
| 基于预训练范式 | 文献[35] | GPT-2 | ATIS、SNIPS | 生成的样本语义可控性较强 | 样本的语义多样性较弱 |
| | 文献[36] | mBERT | Spanish、Thai | 动态增强语义多样性 | 模型规模较大,训练耗时 |
| | 文献[37] | GPT-3 | HWU64、CLINC150 | 结合上下文语义生成意图样本 | 生成样本的质量不稳定 |
| | ICDA[38] | RoBERTa | BANKING、HWU、CLINC | 生成样本的质量可控性较强 | 生成的样本质量易受到模型规模的限制 |

## 2.4 基于度量学习的小样本口语理解

目前,基于度量学习的方法已经成为解决小样本口语理解任务的主流方法,如图5所示,其主要思想是利用度量函数计算两个样本之间的距离,从而得到它们之间的相似度。

### 2.4.1 原型网络

2017年,Snell等[39]为解决小样本分类问题提出原型网络(prototypical networks,PN)如图6所示。该模型的整体思想是首先通过学习一个度量空间,

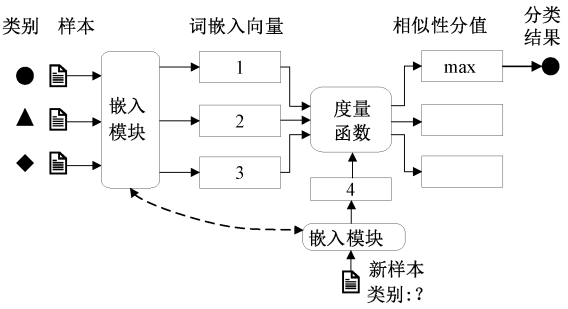


**图5 基于度量学习的口语理解示意图**

**Figure 5 Schematic of SLU with metric learning**

在该空间中用每一类样本的平均值作为该类别的样本中心,对于查询集新样本 $x$,计算 $x$ 与每一类样本中心的欧氏距离,选择距离最小的类作为查询集新样本 $x$ 的最终分类。与其他的小样本学习方法相比,该模型的分类器具有较强的泛化性,同时使用样本中心表示类别,提高了模型的鲁棒性。

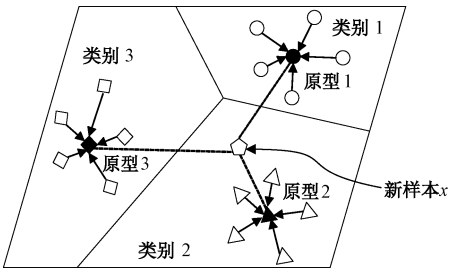


**图 6　原型网络示意图**

**Figure 6　Schematic of the prototypical network**

2020 年,Hou 等[40]在原型网络基础上,设计出基于相似性度量的小样本学习模型 SepProto,以及利用 Goo 等[14]提出的门控机制设计出 JointProto 模型,实现意图分类和槽位填充的联合学习。利用对话意图分类领域新的研究基准 FewJoint 在两个基于原型网络的模型上进行实验,结果表明:JointProto 模型在意图分类和槽位填充两个任务上都优于 SepProto,前者意图分类的准确率高于后者 7.25%,证明了来自联合学习任务的额外信息能够提高模型的性能,与普通的小样本学习方法相比,联合学习在语言理解上更具有优势。2021 年,Xu 等[41]提出语义传输原型网络(semantic transportation prototypical network, STPN),是首个专注于单词级判别信息的小样本意图分类模型。Xu 等[41]认为在度量空间中,不相关的词会导致同一类词的全局特征表示相距较远。2021 年,Dopierre 等[42]在原型网络的基础上进行扩展,提出了一种应用在意图分类任务中的短文本分类元学习算法 PROTAUGMENT。Dopierre 等[42]认为元学习模型在小样本训练过程中很容易导致过拟合,通过在原型网络框架中引入一种无监督离散释义损失去解决该问题。将自动编码器在 Seq2Seq 任务上进行预训练的去噪过程转化为释义生成任务,不同的解码方法大多使用基于 Beam Search 算法进行扩展,使用 Diverse Beam Search (DBS)算法替代 Beam Search 算法,进一步提高了释义的多样性。2022 年,Yang 等[43]认为由于训练数据有限,难以覆盖用户的多样性表达,导致如今的小样本学习方法在小样本口语理解任务中效果较差。受到 Word2Vec 模型中单词类比关系的启发,提出了一种多样性特征增强的原型网络(diversity features enhanced prototypical network, DFEPN)模型,通过对已知意图样本的多样性特征进行充分挖掘,并将其迁移到新的意图样本中,从而达到增强新的意图样本多样性特征的效果。

### 2.4.2　归纳网络

基于度量学习的神经网络架构往往致力于将新的查询集样本与支持集中的样本进行比较,2019 年,Geng 等[44]认为同一类别的不同表述有很多种,这种比较会忽视从样本表示到类别表示的建模。因此在 Yang 等[43]的启发下,将小样本学习方法和胶囊网络进行融合,提出了一种新颖的归纳网络(induction networks,IN)。使用胶囊和动态路由从基于样本的广义类级表示中捕获信息,动态路由方法使模型在小样本文本分类任务中具有更好的泛化能力。模型采用了 Encoder-Induction-Relation 三级框架,架构如图 7 所示。其中 Encoder 模块使用基于自注意力机制的 BiLSTM 编码输入的词向量矩阵,得到每个样本的句子级别语义表示;支持集中每个样本被编码为样本向量后,Induction 模块将其视为胶囊的输入,经过 Dynamic Routing 变换后,输出胶囊归纳出支持集样本的类别特征;Relation 模块用于度量查询集和类别之间的语义关系,进而完成分类。

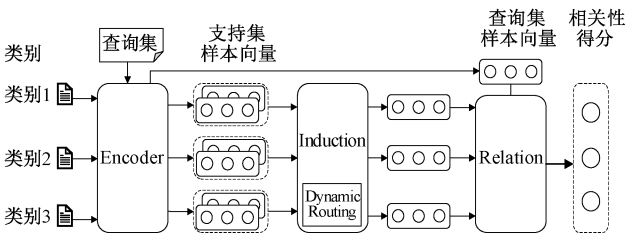


**图 7　归纳网络架构图**

**Figure 7　Induction networks architecture diagram**

2020 年,Geng 等[45]在 IN 的基础上进行改进,提出动态记忆归纳网络(dynamic memory induction networks, DMIN),与文献[44]区别在于编码模块采用 BERT-base,并增加了预训练监督学习阶段。经过动态记忆模块后,样本向量得到更好的分离,动态记忆模块能够有效利用监督学习的经验来编码低级别样本特征和高级别样本特征之间的语义关系,从而实现小样本文本分类。

在小样本学习方法中,相似性度量模型除了原型网络、归纳网络,还有孪生网络、匹配网络[46]和关系网络[47]等,但针对后三者的研究主要集中在图像领域。如表 5 所示,在小样本口语理解任务中,将基于相似性度量的方法与深度学习方法相结合,是该领域今后的研究重点。

表5 基于度量学习的小样本口语理解模型对比

Table 5 Comparing few-shot spoken language understanding models with metric learning

| 分类 | 模型 | 基础模型 | 数据集 | 贡献 | 缺陷 |
|---|---|---|---|---|---|
| 原型网络 | PN[39] | Proto Net | Omniglot、miniImageNet | 经典解决小样本分类问题的方法 | 易受标注数据偏差的影响 |
| | JointProto[40] | Proto Net | FewJoint、SMP2020 | 第一个针对联合多任务学习的小样本NLP基准 | 静态度量方法,未考虑不同样本权重的影响 |
| | STPN[41] | Proto Net | NLUE | 首个聚焦单词级判别信息的小样本口语理解模型 | 捕获样本多样性特征能力较弱 |
| | PROTAUGMENT[42] | Proto Net | BANKING77、HWU64 | 利用无监督方法解决小样本学习过拟合问题。 | 易受到样本中噪声的影响 |
| | DFEPN[43] | Proto Net | SNIPS、NLUE | 提高了样本特征多样性 | 泛化能力较弱 |
| 归纳网络 | IN[44] | IN | ARSC、ODIC | 首次将意图分类任务转化为NLI任务,鲁棒性好 | 无法量化NLI在意图分类任务中的性能 |
| | DMIN[45] | BERT-base | miniRCV1、ODIC | 利用预训练监督学习的知识分离样本 | 模型规模较大,训练耗时 |

## 3 挑战与前沿

随着对话式人工智能的持续发展,新的口语理解任务不断出现。在实际的应用场景中,用户表达的语义具有多样性,目前的预训练语言模型并不能真正解决深度学习模型鲁棒性差、可解释性弱、推理能力缺失等问题。

### 3.1 零样本口语理解

零样本口语理解的任务是使模型能够在没有接受样本训练的情况下,对用户输入的内容进行识别和理解。目前针对零样本口语理解任务可以从以下3个方面进行研究:①借助外部资源,将现有意图中的先验知识转移到新意图中,从而实现对新意图的推断预测。但该方法需要对每一种新意图添加额外的辅助信息,代价十分昂贵;②基于相似性学习的方法度量新意图标签和已知意图样本之间的相似性,但在不同的语境中,语义会发生动态变化,从而产生语义漂移问题;③利用槽位填充任务指导意图分类,两个任务联合建模有助于提高意图分类的准确率。但两个任务产生的噪声会在模型中传播,如何有效控制噪声、对有用知识进行增强,是未来的主要研究方向之一。

### 3.2 中文口语理解

目前,针对中文的口语理解研究远不如对英文的口语理解研究,其中一方面的原因是带有标注的中文意图训练数据较少,对中文文本进行标注代价十分昂贵;另一方面是中文具有比英文更为复杂的结构,表达的语义更加丰富。

2021年,Sun等[48]提出ERNIE 3.0模型在各种NLP任务中表现出比已有的中文预训练语言模型更加出色的效果。ERNIE 3.0模型的参数量更少,在小样本环境中可以快速进行模型微调和训练,同时在情感分析、口语理解等任务上表现出强大的性能。但在中文意图识别任务中,ERNIE模型的潜力还有进一步挖掘的空间,是未来该领域的工作者进一步研究的方向之一。

### 3.3 开放域口语理解

现阶段的对话系统大多停留在封闭的知识领域内,在真实的应用场景中,更多的是需要解决开放领域的问题。2022年,Zhang等[49]构建了两个用于开放域意图分类的数据集CLINC-Single-Domain-OOS与BANKING77-OOS。作者在BERT模型上进行验证后发现,经过预训练的Transformer模型在两个数据集上的鲁棒性很差,开放域场景下的小样本口语理解还需要进行细粒度更好的研究。

目前针对开放域意图识别可以细分为两个子任务:①将开放域意图与已知域内意图分离;②捕获开放域意图的细粒度类别。模型需要在确保“已知意图”准确识别的前提下,捕获没有先验知识的“未知意图”。未来的研究需要寻找合适的决策边界,平衡对“已知意图”和“未知意图”的识别能力。

### 3.4 跨语言口语理解

目前大多数针对口语意图识别的研究以英文为主。但对于不流行或者资源较少的语言来说,在口语理解任务中,同样需要找到一种合适的解决方案。现阶段针对跨语言口语理解的研究主要有两种方式。一种是选择基准英语数据集,将其翻译成目标语言。2020年,Bhathiya等[50]先在英语样本中学习先验知识,接着在西班牙语和泰语样本上验证模型的适应性。该方法存在的缺陷是机器翻译时会出现

数据扭曲问题,样本质量显著降低而无法训练语言模型。另一种是利用迁移学习的方法。2021 年,Sharma 等[51]提出多语言教师-学生网络(multi-lingual teacher-student network,MTSN),将从 mBERT 模型中学习到的先验知识迁移到目标语言任务中。该方法减少了对目标语言样本量的需求,但会受到不同语种表达方式的差异而显著影响模型的性能。

## 4　结束语

在口语理解领域中,基于大数据的预训练语言模型已经在传统的口语理解数据集上取得了接近饱和的效果。相比研究经典数据集,在现实的应用场景中,更多面对的是训练样本受限的问题。近年来,随着小样本学习方法在图像领域的深入研究,越来越多的 NLP 领域研究者开始关注该方法在口语理解任务中的应用。本文重点阐述了小样本场景下的模型微调、数据增强和度量学习 3 类方法,对不同模型的可解释性、推理能力以及泛化能力等性能进行对比。未来的研究重点是用户在不同场景下语义多样性的表达,以进一步提高模型在深层次语义上的理解能力。

## 参考文献:


[1] 薛均晓, 黄世博, 王亚博, 等. 基于时空特征的语音情感识别模型 TSTNet[J]. 郑州大学学报(工学版), 2021, 42(6): 28-33.
XUE J X, HUANG S B, WANG Y B, et al. Speech emotion recognition TSTNet based on spatial-temporal features[J]. Journal of Zhengzhou University (Engineering Science), 2021, 42(6): 28-33.

[2] LOUVAN S, MAGNINI B. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: a survey[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: ICCI, 2020: 480-496.

[3] WELD H, HUANG X Q, LONG S Q, et al. A survey of joint intent detection and slot filling models in natural language understanding[J]. ACM Computing Surveys, 2022, 55(8): 1-38.

[4] TUR G, DE MORI R. Spoken language understanding: systems for extracting semantic information from speech[D]. New York: John Wiley and Sons, 2011.

[5] BHARGAVA A, CELIKYILMAZ A, HAKKANI-TÜR D, et al. Easy contextual intent prediction and slot detection[C]//International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2013: 8337-8341.

[6] MESNIL G, DAUPHIN Y, YAO K S, et al. Using recurrent neural networks for slot filling in spoken language understanding[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(3): 530-539.

[7] LIN Z H, FENG M W, DOS SANTOS C N, et al. A structured self-attentive sentence embedding[EB/OL]. (2017-03-09)[2023-08-09]. http://export.arxiv.org/abs/1703.03130.

[8] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. (2014-09-03)[2023-08-09]. https://arxiv.org/abs/1408.5882.

[9] ZHAO W, YE J B, YANG M, et al. Investigating capsule networks with dynamic routing for text classification[EB/OL]. (2018-09-03)[2023-08-09]. https://arxiv.org/abs/1804.00538.

[10] ZHANG X D, WANG H F. A joint model of intent determination and slot filling for spoken language understanding[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: ACM, 2016: 2993-2999.

[11] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. (2014-12-11)[2023-08-09]. https://arxiv.org/abs/1412.3555.

[12] LIU B, LANE I. Joint online spoken language understanding and language modeling with recurrent neural networks[EB/OL]. (2016-09-06)[2023-08-09]. https://arxiv.org/abs/1609.01462.

[13] LIU B, LANE I. Attention-based recurrent neural network models for joint intent detection and slot filling[EB/OL]. (2016-09-06)[2023-08-09]. https://arxiv.org/abs/1609.01454.

[14] GOO C W, GAO G, HSU Y K. Slot-gated modeling for joint slot filling and intent prediction[C]// The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Haven: ACL, 2018:753-757.

[15] QIN L B, CHE W X, LI Y M, et al. A stack-propagation framework with token-level intent detection for spoken language understanding[EB/OL]. (2019-09-05)[2023-08-09]. https://arxiv.org/abs/1909.02188.

[16] CHEN M Y, ZENG J, LOU J. A self-attention joint model for spoken language understanding in situational dialog applications[EB/OL]. (2019-05-27)[2023-08-09]. https://arxiv.org/abs/1905.11393.

[17] WANG Y, SHEN Y L, JIN H X. A Bi-model based RNN semantic frame parsing model for intent detection and slot filling[EB/OL]. (2018-12-26)[2023-08-09]. https://arxiv.org/abs/1812.10235.

[18] CHEN Q, ZHUO Z, WANG W. BERT for joint intent classification and slot filling[EB/OL]. (2019-02-28)[2023-08-09]. https://arxiv.org/abs/1902.10909.

[19] QIN L B, LI Z Y, CHE W X, et al. Co-GAT: a co-interactive graph attention network for joint dialog act recognition and sentiment classification[EB/OL]. (2020-12-24)[2023-08-09]. https://arxiv.org/abs/2012.13260.

[20] LI F F, FERGUS R, PERONA P. One-shot learning of object categories[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4): 594-611.

[21] KOCH G, ZEMEL R, SALAKHUTDINOV R. Siamese neural networks for one-shot image recognition[C]// International Conference on Machine Learning. Piscataway: IEEE, 2015: 1-30.

[22] ZHANG X T, QIANG Y T, SUNG F, et al. RelationNet2: deep comparison columns for few-shot learning[EB/OL]. (2018-11-17)[2023-08-09]. https://arxiv.org/abs/1811.07100.

[23] CHEN J F, ZHANG R C, MAO Y Y, et al. ContrastNet: a contrastive learning framework for few-shot text classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10492-10500.

[24] JIAN Y R, TORRESANI L. Label hallucination for few-shot classification[EB/OL]. (2021-12-06)[2023-08-09]. https://arxiv.org/abs/2112.03340.

[25] DAI A M, LE Q V. Semi-supervised sequence learning[EB/OL]. (2015-11-04)[2023-08-09]. https://arxiv.org/abs/1511.01432.

[26] HOWARD J, RUDER S. Universal language model fine-tuning for text classification[EB/OL]. (2018-05-23)[2023-08-09]. https://arxiv.org/abs/1801.06146.

[27] SUN C, QIU X P, XU Y G, et al. How to fine-tune BERT for text classification? [J]. Lecture Notes in Computer Science, 2019, 11856: 194-206.

[28] MOHAMMADI S, CHAPON M. Investigating the performance of fine-tuned text classification models based-on Bert[C]//2020 IEEE 22nd International Conference on High Performance Computing and Communications. Piscataway: IEEE, 2020: 1252-1257.

[29] ZHANG H D, ZHANG Y W, ZHAN L M, et al. Effectiveness of pre-training for few-shot intent classification[EB/OL]. (2021-09-13)[2023-08-09]. https://arxiv.org/abs/2109.05782.

[30] ZHANG H D, LIANG H W, ZHANG Y W, et al. Fine-tuning pre-trained language models for few-shot intent detection: supervised pre-training and isotropization[EB/OL]. (2022-05-26)[2023-08-09]. https://arxiv.org/abs/2205.07208.

[31] KURATA G, XIANG B, ZHOU B W, et al. Labeled data generation with encoder-decoder LSTM for semantic slot filling[C]//17th Annual Conference of the International Speech Communication Association. San Francisco: ISCA, 2016: 725-729.

[32] HOU Y T, LIU Y J, CHE W X, et al. Sequence-to-sequence data augmentation for dialogue language understanding[EB/OL]. (2018-06-04)[2023-08-09]. https://arxiv.org/abs/1807.01554.

[33] KIM H Y, ROH Y H, KIM Y K. Data augmentation by data noising for open-vocabulary slots in spoken language understanding[C]// The 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). New Haven: ACL, 2019: 97-102.

[34] ZHAO Z J, ZHU S, YU K. Data augmentation with atomic templates for spoken language understanding[EB/OL]. (2019-08-28)[2023-08-09]. https://arxiv.org/abs/1908.10770.

[35] PENG B L, ZHU C G, ZENG M, et al. Data augmentation for spoken language understanding via pretrained language models[EB/OL]. (2021-03-11)[2023-08-09]. https://arxiv.org/abs/2004.13952.

[36] QIN L B, NI M H, ZHANG Y, et al. CoSDA-ML: multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP[EB/OL]. (2020-07-13)[2023-08-09]. https://arxiv.org/abs/2006.06402.

[37] SAHU G, RODRIGUEZ P, LARADJI I H, et al. Data augmentation for intent classification with off-the-shelf large language models[EB/OL]. (2022-04-05)[2023-08-09]. https://arxiv.org/abs/2204.01959v1.

[38] LIN Y T, PAPANGELIS A, KIM S, et al. Selective in-context data augmentation for intent detection using pointwise V-information[EB/OL]. (2023-02-10)[2023-08-09]. https://arxiv.org/abs/2302.05096v1.

[39] SNELL J, SWERSKY K, ZEMEL R S. Prototypical networks for few-shot learning[EB/OL]. (2017-06-19)[2023-08-09]. https://arxiv.org/abs/1703.05175.

[40] HOU Y T, MAO J F, LAI Y K, et al. FewJoint: a few-shot learning benchmark for joint language understanding[EB/OL]. (2020-12-13)[2023-08-09]. https://arxiv.org/abs/2009.08138.

[41] XU W Y, ZHOU P L, YOU C Y, et al. Semantic transportation prototypical network for few-shot intent detection[C]//Interspeech 2021. Brno, Czechia: ISCA, 2021: 251-255.

[42] DOPIERRE T, GRAVIER C, LOGERAIS W. PROTAUGMENT: unsupervised diverse short-texts paraphrasing for intent detection meta-learning[EB/OL]. (2021-05-27)

[2023-08-09]. https://arxiv.org/abs/2105.12995.

[43] YANG F Y, ZHOU X, WANG Y, et al. Diversity features enhanced prototypical network for few-shot intent detection [C]// International Joint Conference on Artificial Intelligence. Vienna, Austria: IJCAI, 2022: 4447-4453.

[44] GENG R Y, LI B H, LI Y B, et al. Induction networks for few-shot text classification[EB/OL]. (2019-09-29) [2023-08-09]. https://arxiv.org/abs/1902.10482.

[45] GENG R Y, LI B H, LI Y B, et al. Dynamic memory induction networks for few-shot text classification [EB/OL]. (2020-05-12)[2023-08-09]. https://arxiv.org/abs/2005.05727.

[46] VINYALS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning[EB/OL]. (2017-12-29) [2023-08-09]. https://arxiv.org/abs/1606.04080.

[47] SUNG F, YANG Y X, ZHANG L, et al. Learning to compare: relation network for few-shot learning [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1199-1208.

[48] SUN Y, WANG S H, FENG S K, et al. ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation[EB/OL]. (2021-07-05) [2023-08-09]. https://arxiv.org/abs/2107.02137.

[49] ZHANG J G, HASHIMOTO K, WAN Y, et al. Are pre-trained transformers robust in intent classification? a missing ingredient in evaluation of out-of-scope intent detection[EB/OL]. (2022-04-07)[2023-08-09]. https://arxiv.org/abs/2106.04564.

[50] BHATHIYA H S, THAYASIVAM U. Meta learning for few-shot joint intent detection and slot-filling[C]//ICMLT 2020: 2020 5th International Conference on Machine Learning Technologies. New York: ACM, 2020: 86-92.

[51] SHARMA B, MADHAVI M, ZHOU X H, et al. Exploring teacher-student learning approach for multi-lingual speech-to-intent classification[C]//2021 IEEE Automatic Speech Recognition and Understanding Workshop. Piscataway: IEEE, 2022: 419-426.

## A Survey of Spoken Language Understanding Based on Few-shot Learning

LIU Na[1,2], ZHENG Guofeng[1,2], XU Zhenshun[1,2], LIN Lingde[1,2], LI Chen[1,2], YANG Jie[1,2]

(1. School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China; 2. The Key Laboratory of Images and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China)

**Abstract**: Few-shot spoken language understanding (SLU) is one of the urgent problems in dialogue artificial intelligence (DAI). The relevant literature on SLU task, combining the latest research trends both domestic and foreign was systematically reviewed. The classic methods for SLU task modeling in non-few-shot scenarios were briefly introduced, including single modeling, implicit joint modeling, explicit joint modeling, and pre-trained paradigms. The latest studies in few-shot SLU were introduced, which included three kinds of few-shot learning methods based on model fine-tuning, data augmentation and metric learning. Representative models such as ULMFiT, prototypical network, and induction network were discussed. On this basis, the semantic understanding ability, interpretability, generalization ability and other performances of different methods were analyzed and compared. Finally, the challenges and future development directions of SLU tasks were discussed, it was pointed out that zero-shot SLU, Chinese SLU, open-domain SLU, and cross-lingual SLU would be the research difficulties in this field.

**Keywords**: spoken language understanding; few-shot learning; fine-tune; data augmentation; metric learning