南京大学

研 究 生 毕 业 论 文

（申请博士学位）

论 文 题 目　面向任务型对话系统的分布外检测方法研究

作 者 姓 名　欧阳亚文

专 业 名 称　计算机科学与技术

研 究 方 向　自然语言处理

指 导 教 师　戴新宇 教授

2023 年 8 月 16 日

学　　号：DZ1833020

论文答辩日期：2023 年 8 月 16 日

指 导 教 师：　　（签字）

Research on Out-of-distribution Detection Methods for Task-oriented Dialogue Systems

by

Yawen Ouyang

Supervised by

Xinyu Dai,Professor

A dissertation submitted to

the graduate school of Nanjing University

in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer science and technology

Computer science and technology

Nanjing University

August 16, 2023

南京大学学位论文原创性声明

本人郑重声明，所提交的学位论文是本人在导师指导下独立进行科学研究工作所取得的成果。除本论文中已经注明引用的内容外，本论文不包含其他个人或集体已经发表或撰写过的研究成果，也不包含为获得南京大学或其他教育机构的学位证书而使用过的材料。对本文的研究做出重要贡献的个人和集体，均已在论文的致谢部分明确标明。本人郑重申明愿承担本声明的法律责任。

研究生签名：_____

日期：_____

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：面向任务型对话系统的分布外检测方法研究

计算机科学与技术 专业 2018 级博士生姓名：欧阳亚文

指导教师（姓名、职称）：戴新宇 教授

摘 要

任务型对话系统通过对话的形式，充当用户助手，给用户提供特定的服务，是近年来自然语言处理领域热门的研究方向。该系统中的意图理解模块旨在理解用户的输入，是后续提供服务的基础。目前，既有意图理解方法大多针对封闭环境，会假设训练和测试阶段的用户输入符合独立同分布，然而当系统被开发完成上线后，其面临的环境往往是开放的，因此，系统会面临训练集分布外的用户输入，然而，既有方法往往仅关注分布内输入的分类性能，难以检测包含未知意图的分布外用户输入。为了解决这一问题，本文针对不同场景下的任务型对话系统，提出了相应的检测方法，主要工作如下：

单标注场景下的分布外检测。针对用户输入仅包含单个意图的场景，本文提出了基于分布外数据增广的校准策略来提升检测性能。主流方法仅依赖于分布内数据来优化检测器，这会使得检测器受到句中偏差词的影响，误把分布外样本分类为分布内样本。为解决这一问题，本文先提出了基于梯度的偏差词检测方法，接着从外部的公开数据集中，筛选带有相同偏差词的分布外样本来校准检测器。实验表明，本文提出的偏差词检测方法可以准确地检测偏差词，检索的分布外样本可以有效地校准检测器，提升其对分布外用户输入的检测能力。

多标注场景下的分布外检测。针对用户输入包含单个或多个意图的场景，本文提出了基于意图数目协助的方案来检测分布外用户输入。既往研究多聚焦于单标注场景，本文首次把该任务扩展到了多标注场景，并指明该场景下的新的挑战：检测同时包含已知和未知意图的混合分布外用户输入。为了应对这一挑战，本文提出基于意图数目协助的方案，通过预测用户输入的意图数，并评估其是否包含相同数量的已知意图，进而检测其是否包含未知意图。实验表明，与已有的单标注方案相比，该方案更适合多标注场景，能取得超过 10% 的显著提升，

尤其是对于混合分布外用户输入的检测。

无标注场景下的分布外检测。针对无意图标注的场景，本文提出了基于检索增强和似然比的方案来检测分布外用户输入。主流检测器的优化过程往往是有监督的，本文方案可以适用于无监督场景，即检测器的训练无需依赖意图标签。本文先从模型层面和数据层面分析了现有基于似然检测的问题：生成式模型对分布内文本的建模效果不佳和文本边缘似然的偏差，在此基础上本文通过引入 k 近邻检索和预训练模型校准的方式，来提升建模效果，缓解文本边缘似然偏差的影响。实验表明，该方案可以有效地解决上述问题，进而比已有的无监督方案取得性能上超过 10% 的显著提升。

计算资源受限场景下的分布外检测。以上场景都是针对计算资源充足的场景，但考虑到预训练语言模型已经成为对话系统的基石，针对存储调优后的预训练语言模型所需开销较大的问题，本文提出了无需调优的轻量方法来检测分布外输入，以适用于计算资源受限场景。具体而言，本文提出了一种基于前缀调优的分布外检测框架，在训练时不需要调优预训练模型，从而不需要存储调优后的模型，大大节省资源开销。同时，本文提出的框架可以适用于无标注、有标注，以及分布外样本参与训练的场景。实验表明，本文方法在轻量化的同时，可以获得相比调优范式有竞争力的结果。

关键词：任务型对话系统；分布外检测；开放环境；意图理解

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Research on Out-of-distribution Detection Methods for Task-oriented Dialogue Systems

SPECIALIZATION: Computer science and technology

POSTGRADUATE: Yawen Ouyang

MENTOR: Xinyu Dai,Professor

ABSTRACT

Task-oriented dialogue systems, which serve as user assistants and provide specific services through conversational interactions, have become a popular research direction in the field of natural language processing in recent years. The intent understanding module within these systems aims to comprehend user inputs and serves as the foundation for subsequent service provision. Currently, existing intent understanding methods mostly target closed environments, assuming that user inputs during training and testing phases follow the same distribution. However, once the system is developed and deployed, it often faces an open environment, where it encounters user inputs outside the training distribution. Existing methods primarily focus on the classification performance of in-distribution inputs, making it challenging to detect out-of-distribution inputs containing unknown intents. To address this issue, this paper proposes detection methods tailored to various scenarios of task-oriented dialogue systems. The main contributions are as follows:

Out-of-Distribution Detection in Single-Label Scenarios: For scenarios where user inputs contain only a single intent, this paper proposes a calibration strategy based on out-of-distribution data augmentation to enhance detection performance. Conventional methods rely solely on in-distribution data to optimize detectors, which can lead to misclassification of out-of-distribution samples due to biased words within sentences. To overcome this, the paper introduces a gradient-based biased word detection method and then calibrates the detector using out-of-distribution samples with similar biased words from external publicly available datasets. Experimental results show that the

proposed biased word detection method accurately identifies biased words and the retrieved out-of-distribution samples effectively calibrate the detector, enhancing its capability to detect out-of-distribution user inputs.

Out-of-Distribution Detection in Multi-Label Scenarios: For scenarios where user inputs contain one or more intents, the paper proposes an intent-number-assisted approach to detect out-of-distribution user inputs. Previous research has mainly focused on single-label scenarios, but this paper extends the task to multi-label scenarios and identifies new challenges in such scenarios—detecting a mixture of known and unknown intents within out-of-distribution user inputs. To address this challenge, the paper introduces an intent-number-assisted approach that predicts the number of intents in user inputs and assesses whether they include the same number of known intents, thereby detecting unknown intents. Experimental results demonstrate that this approach is better suited for multi-label scenarios and achieves significant performance improvements of over 10%, especially for detecting mixed out-of-distribution user inputs.

Out-of-Distribution Detection in Unlabeled Scenarios: For scenarios without labeled intents, the paper proposes a retrieval-enhanced and likelihood-ratio-based approach to detect out-of-distribution user inputs. Conventional detectors are often optimized in a supervised manner, but this paper's approach is applicable to unsupervised scenarios, meaning the detector's training does not depend on intent labels. The paper analyzes existing likelihood-based detection problems from both model and data perspectives, including poor modeling of in-distribution text by generative models and biases in text marginal likelihood. Based on this analysis, the paper introduces a solution that leverages k-nearest neighbors retrieval and pre-trained model calibration to improve modeling and alleviate the impact of text marginal likelihood bias. Experimental results show that this approach effectively addresses the aforementioned issues and achieves significant performance improvements of over 10% compared to existing unsupervised methods.

Out-of-Distribution Detection in Resource-Constrained Scenarios: The aforementioned scenarios assume sufficient computational resources. However, considering that pre-trained language models have become fundamental to dialogue systems and

the substantial overhead associated with fine-tuning stored pre-trained models, the paper proposes a lightweight method that does not require fine-tuning to detect out-of-distribution inputs, suitable for resource-constrained scenarios. Specifically, the paper presents a prefix-tuning-based out-of-distribution detection framework that eliminates the need for fine-tuning pre-trained models during training, thereby avoiding the storage overhead of fine-tuned models and significantly reducing resource costs. Additionally, the framework is applicable to scenarios involving unlabeled, labeled, and out-of-distribution samples participating in training. Experimental results demonstrate that the proposed method achieves competitive results compared to the fine-tuning paradigm while being lightweight.

KEYWORDS: Task-oriented dialogue system; Out-of-distribution detection; Open environment; Intent understanding

目　　录

中文摘要 I

ABSTRACT III

目　　录 VII

第一章 绪论 1

1.1 引言 1

1.2 本文工作 5

1.3 本文结构 7

第二章 研究现状 9

2.1 任务型对话系统 9

2.2 意图理解 10

2.3 分布外检测 11

2.4 本章小结 19

第三章 单标注场景下的分布外检测 21

3.1 引言 21

3.2 相关工作 23

3.3 本文方法 24

3.4 实验设置 27

3.5 结果和分析 28

3.6 本章小结 33

第四章 多标注场景下的分布外检测 35

4.1 引言 35

4.2 本文方法 37
4.3 实验设置 41
4.4 结果和分析 43
4.5 总结 49

第五章 无标注场景下的分布外检测 51
5.1 引言 51
5.2 相关工作 52
5.3 本文方法 53
5.4 实验设置 57
5.5 结果和分析 59
5.6 总结 61

第六章 计算资源受限场景下的分布外检测 65
6.1 引言 65
6.2 相关工作 67
6.3 本文方法 68
6.4 实验设置 72
6.5 结果和分析 74
6.6 总结 80

第七章 总结与展望 81
7.1 本文贡献 81
7.2 未来方向 82

参考文献 85

致　谢 103

科研成果与学术活动 105

学位论文出版授权书 107

第一章　绪论

1.1　引言

机器能否自然流畅地与人进行对话，是衡量其智能程度的重要标准之一。著名的计算机科学家和数学家艾伦·麦席森·图灵于 1950 年提出图灵测试来评估机器的智能水平，图灵测试的基本流程是，一个人类测试者分别与一个机器和一个人类通过打字机交互，如果超过 30% 的测试者在 5 分钟之内无法分辨机器与人类的回答，则认为机器表现出了人类智能的水平。因此，提升机器与人的对话能力，构建切实可行的对话系统对推动人工智能领域发展具有极为重要的科学意义。

除了在科学研究领域中的重要性，对话系统在现实生活中亦有着广泛的落地场景。首先，政府可以通过对话系统为民众提供政务办事的引导，例如在各级政府机关网站上设置智能问答系统，让民众能够方便快捷地获取所需信息，从而提升民众的办事效率，并减轻公务人员的负担。在疫情期间，对话系统在为公众提供防疫信息的同时，避免了人与人的直接接触，可以减缓病毒的传播。其次，对企业而言，构建 24 小时在线的客服机器人可以提高企业的服务效率和质量。这种机器人可以随时回答用户的问题，为用户提供快速的解决方案，不受时间和空间的限制，现如今，淘宝、京东等电商公司已经广泛使用这种技术。最后，对个人而言，对话系统不仅可以充当用户助手，帮助用户高效地完成订餐，订火车票等服务，还可给予用户全天不间断的情感陪护。对于那些需要寻求安慰或支持的人，对话系统可以随时提供陪聊和帮助，无需预约或等待，甚至无需缴费。综上所述，对话系统可以被广泛应用于政府、企业、个人领域，为整个社会都可以带来巨大的便利。

典型的对话系统研究聚焦于任务型（task-oriented）对话系统和闲聊型（chit-chat）对话系统[1]，具体例子可参考图 1-1，相较于闲聊型，任务型对话系统需要精准理解用户输入：将用户输入转化成系统能够理解和处理的结构化形式，意

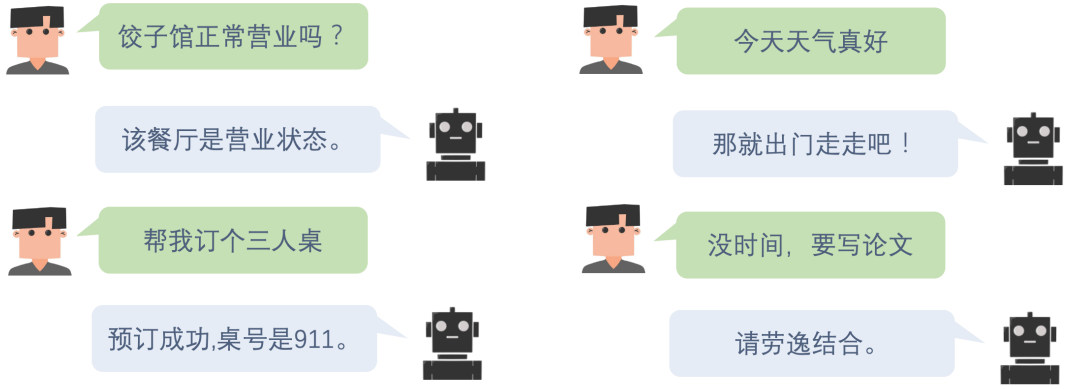

图 1-1　任务型对话系统（左图）和闲聊型对话系统（右图）。

图理解是其中的关键步骤。任务型对话系统旨在帮助用户完成特定的服务，在此过程中，精准识别用户意图，可保证对话过程的高效，从而节省用户时间与精力。相反，闲聊型对话系统旨在与用户进行无目的交流，没有任务限制，该类系统通常不需要精准理解用户输入，而是旨在为用户提供愉快的对话体验。

意图理解可形式化为文本分类任务，输入为文本形式的用户话语（utterance），输出为来自预定义意图空间的意图标签。举例说明，预定义的意图空间可包括“订车票”、“查天气”、“查快递”等，用户话语“查查南京今天的天气”的意图为“查天气”。为了准确地进行预测，早期的研究人员采用基于规则的方式进行意图理解。这种方法主要是基于预定义的规则，比如，一旦检测到用户话语中包含特定的关键字，就输出相应的意图，举例说明，对于上述的用户输入，模型一旦检测到其中包含“天气”这个关键字，就输出“查天气”的意图。然而，这种基于规则的方法泛化能力较弱，而且需要大量的人工设计规则，无法应对真实世界中，用户表达的多样化和复杂化。后续的研究者通常采用经典的机器学习的方法，如支持向量机（support vector machine，简称 SVM）等，该类方法的泛化性强于基于规则的方式，但往往伴随着复杂的特征工程，需要耗费大量的时间和资源。随着深度学习的流行，近期的工作采用基于神经网络的方法，如基于 Transformer 的文本分类[2]，这些基于神经网络的方法不仅具有较强的泛化性，同时可以减少人工预处理，更适合于处理复杂的自然语言输入，成为近些年来意图理解的主要方法[3-5]。

尽管基于神经网络的方法在对话理解上取得了不俗的成功，目前主流的研究工作还是基于封闭世界的假设。即假设当任务型对话系统开发完成上线后，用

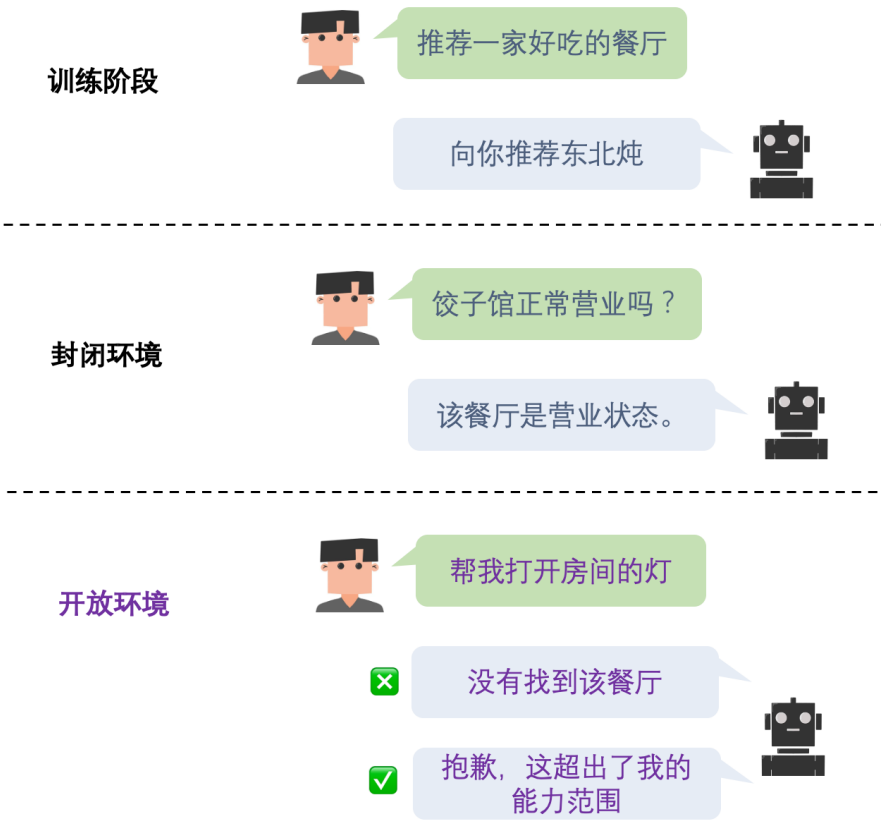

图 1-2 开放环境下分布外用户输入的例子

户输入的分布依旧和训练集相同，不会出现来自预定义意图空间外的未知意图。举例说明，对于一个专门为订餐领域设计的对话系统，预定义的意图空间都是与订餐相关的，基于封闭世界的对话系统会假设该系统在往后的使用过程中，用户的输入亦都与订餐相关，不考虑其他领域的用户输入，如购物、预订机票等。

然而，当一个机器学习模型开发完成上线后，其往往面临着开放环境[6]，任务型对话系统也不例外，在该环境下，用户的输入可能来自训练集分布外，即表达对话系统未定义的意图。如图 1-2 所示，对于订餐领域的对话系统，用户输入“帮我打开房间的灯”属于分布外，因为其意图并不属于订餐领域，即不存在于预定义的意图集合中。

可靠的任务型对话系统应该检测分布外用户输入，避免出现答非所问。如果对话系统对于分布外输入不做额外检测，直接把它分类到已知类别空间中，就会出现答非所问的情况，浪费用户时间。相反地，如果对话系统能够识别出分布外输入，并输出安全回复，如“抱歉，这超出了我的能力范围”，则可以一定程度上实现“优雅地”失败，提升用户的使用体验。同时，检测到的分布外输入也能为对话系统的后续开发提供指导。

为实现开放环境下可靠的任务型对话系统，本文开展了不同场景下的分布外检测研究。具体而言，本文从单标注、多标注、无标注、计算资源受限四个场景出发，较为系统地为对话系统设计不同的分布外检测方法，最终提升开放环境下意图理解的稳健性（robustness）。单标注场景是假设训练和测试阶段，用户的输入中只包含单个意图，比如，上文所举的例子“查查南京今天的天气”为单标注的用户输入，这一场景虽然简单，但却不一定满足真实场景。多标注场景打破了用户输入只包含单个意图的假设，允许用户输入中可以包含一个或多个意图，比如，“查查南京的天气，并导航去那里”为包含多意图的用户输入，相较于单标注，多标注场景更符合真实场景，且更具有挑战性。无标注场景是在构建分布外检测器时，不使用意图标注，旨在为意图标注昂贵的领域，探索更多可能的技术路线。以上场景都假设存储资源充足的场景，考虑到预训练语言模型和调优范式已经成为分布外检测方法的基石，而为每一场景都存储调优后的模型会非常消耗资源，计算资源受限场景旨在解决这一问题，在构建分布外检测器时，不调优基石模型，在节省成本的同时，实现分布外检测。

本文的研究在科学层面和现实层面都具有重要意义，包括但不限于以下几方面：

- **提升神经网络输出的可靠性**。分布外检测可以帮助神经网络过滤超出其能力范围，即不属于训练类别的样本，保证输入都来自训练类别，进而提升其输出内容的可靠性。
- **避免对话系统答非所问，提升用户使用体验**。当检测到分布外的用户输入后，对话系统可以直接给予用户预定义的安全回复，避免把该输入当成分布内样本处理，进而出现答非所问，误导用户的情况。这不仅可以使得对话系统及时发现自身的不足，亦可节约用户的时间与精力，提高用户对对话系统的满意度和信任度。
- **指导对话系统的后续开发和优化**。收集到的分布外数据对于对话系统的后续开发和优化具有重要的指导意义。对于分布外数据，开发人员可以进行保存和进一步的数据分析，如聚类、可视化等。通过对分布外数据的分析，可以了解目前系统的局限性和不足之处，发掘用户的新需求和问题，并为对话系统的迭代和升级提供宝贵的指导和改进方向。

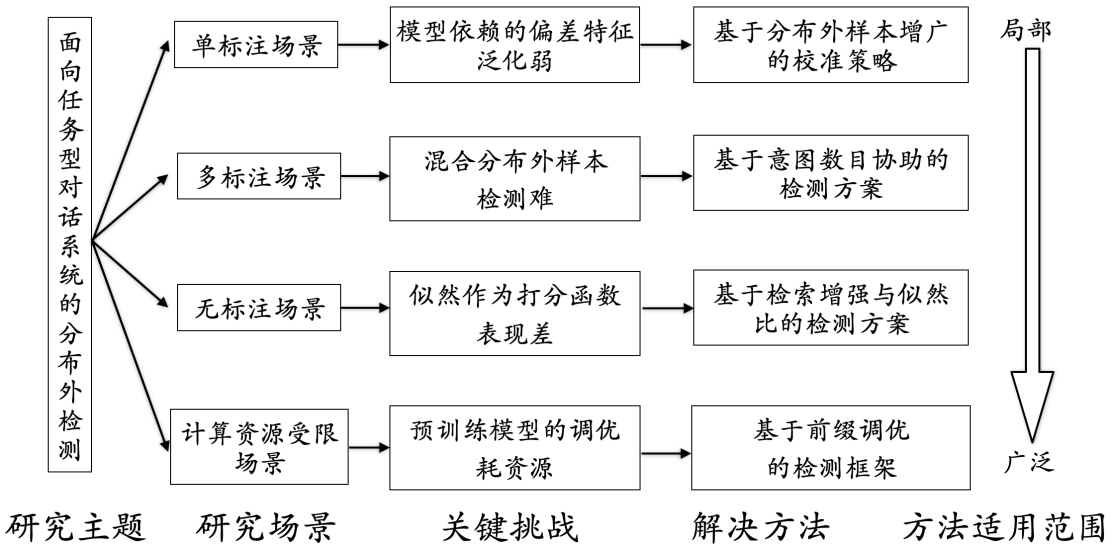

图 1-3 本文的研究内容总览

1.2 本文工作

本文针对开放环境下的意图理解开展研究，分别探索了单标注、多标注、无标注和计算资源受限场景，这些场景都有共同的目标：检测分布外的用户输入，场景不同，关键挑战与解决方案也有不同侧重。总览如图 1-3 所示，具体内容如下：

- **单标注场景分布外检测**。研究者在研究分布外检测任务时，为了简化问题，往往会假设用户输入中只表达单个意图。基于此假设，分布外检测器可基于由训练数据及意图标签，通过交叉熵损失训练得到的已知意图分类器，比如，基于能量的分布外检测器会基于分类器推导出的用户输入的能量值来进行检测：分布内样本的能量值往往较小，分布外样本的能量值往往较大。然而，本文发现，只基于训练数据及标签训练得到的检测器会受到偏差词（biased words）的影响：检测器往往会依赖关注句中的某些偏差词直接作判断，泛化性较弱，例如，对于包含偏差词的分布外句子，受这些词的影响，检测器往往会将其误分类为分布内。为解决该问题，本文提出了基于分布外样本增广的校准方案，具体而言，本文先提出了基于梯度的偏差词检测方法，在检测出偏差词后，基于从外部公开数据集中筛选带有这些偏差词的分布外样本，通过本文引入的新的训练目标，来校准检测器。实验上，本文分别对基于能量和基于马氏距离的检测器进行了校准，结果表明，校准后的检测器可以提升对分布外样本的检测能力。

- **多标注场景分布外检测**。单标注场景虽然较为流行，但真实场景中用户的输入可能包含一个或多个意图，即更符合多标注的假设。关于分布外检测问题，以往的工作往往聚焦于单标注场景，本文首次把单标注扩展到了多标注场景，并且提出了这一场景下的新挑战：对混合分布外用户输入的检测。所谓混合分布外用户输入是指用户输入中既包含已知意图，也包含未知意图。由于单标注场景下用户输入不会包含两个意图，所以混合分布外输入只会在多标注场景中出现。为了应对这一挑战，本文提出了基于意图数目协助的方案，在衡量用户输入是否包含已知意图的同时，也引入预测意图数的任务。训练时，两个任务联合进行训练，测试时，通过衡量用户输入中的意图是否都为已知意图，进而判断用户输入中是否包含未知意图，即用户输入是否为分布外输入。实验表明，本文提出的意图数目协助方案相较于单标注场景下的方案，可以显著提升对分布外用户输入的检测能力，尤其是对于混合分布外用户输入的检测。
- **无标注场景分布外检测**。单标注场景和多标注场景下的检测器都需要利用意图标签，本场景旨在探索不利用意图标注的情况下，实现分布外检测，探寻分布外检测方法的更多可能性。现有的无标注场景下的分布外检测往往基于生成式模型：通过训练生成式模型来拟合训练样本的分布，理想情况下，由该生成器推出的分布外样本的似然较低，分布内样本的似然较高，但实际中，两类样本的似然区分并不明显。在对基于生成式模型的方法进行实验分析和理论推导后，本文发现此类方法在模型层面和数据层面都存在问题：模型对分布内样本建模性能不足和文本边缘似然的偏差。在此基础上，本文通过引入 k 近邻检索的方式来增强模型对分布内样本的建模效果，通过引入预训练语言模型的方式来衡量文本边缘似然，最终提出似然比的方法来缓解文本边缘似然偏差的影响。本文通过贝叶斯定理证明了所提方法的合理性，并在单标注和多标注分布外检测数据集上都进行了测试，实验结果表明，本文提出的方法可以较好地解决上述问题，进而比已有的无监督方法取得性能上的显著提升。
- **计算资源受限场景分布外检测**。以上场景的分布外检测方法都需要基于并调优预训练语言模型，然而，随着预训练语言模型的快速发展，其参数也愈来愈多，为不同数据集、不同场景调优并存储这些模型往往需要较多开销。

针对这一问题，本文首次把分布外检测扩展到了计算资源受限场景，旨在无需调优预训练语言模型的分布外检测。同时，本文为该场景提出了新的框架：通过引入前缀参数，并固定预训练语言模型的参数来拟合训练数据的分布，最终对比有无前缀的用户输入的似然来检测分布外用户输入。在训练前缀参数时，本文提出的方法无需样本标签。同时本文也对该框架进一步扩展，使其在训练前缀时引入标签信息，以及引入可使用的分布外样本。实验表明，本文提出的方法在不调优预训练语言模型的同时，可以获得有竞争力的结果，并大大节省存储模型的开销。引入标签信息和分布外样本后，该方法的性能可以得到进一步提升。除了对话任务，本文也在其他任务的分布外检测上验证了该方法的有效性。

这些场景对实现分布外检测所需的条件逐渐放缓，适用范围也从局部到广泛。单标注场景会假设用户输入中只包含单个标签，多标注场景就打破了这一假设，但这两个场景在训练阶段都需要依赖输入标签，无标注场景打破了这一要求，以上三个场景都需要调优预训练模型，计算资源受限场景可在不依赖标注，同时不调优预训练模型的前提下，实现分布外检测。

作者才疏学浅，文中肯定存在不少错误，遂将会持续更新并维护最新版本，读者可以在https://yawenouyang.github.io/about/files/thesis.pdf上获取，恳请各位读者不吝指正。

1.3 本文结构

本文共包含七章，第一章是绪论，先介绍了对话系统的研究背景与意义，随后总览了本文工作，最后介绍了本文结构。第二章是相关研究，分别介绍了意图理解和分布外检测。第三章是单标注场景下的分布外检测，第四章是多标注场景下的分布外检测，第五章是无标注场景下的分布外检测，第六章是计算资源受限场景下的分布外检测，第七章是总结和对未来研究方向的展望。

第二章　研究现状

2.1　任务型对话系统

任务型对话系统旨在与用户进行自然语言交互，以协助用户完成特定的任务。经典的任务型对话系统在收到用户的输入时，除了进行意图理解，还需要进行对话状态跟踪[7-10]，策略学习[11-14]，对话生成[15-18]，详情见图 2-1 [1]。对于意图理解任务，输入是用户话语，输出是话语意图。对于对话状态跟踪任务，意图理解的输出作为其输入的一部分，并结合用户输入、历史对话状态，共同决定当前轮的对话状态。对话状态由一系列预定义的槽位（地点、口味、时间）与对应的槽值（None、好吃、None）构成。对于策略学习任务，意图理解的输出同样会直接作为其部分输入，输出系统动作（可以理解为系统的意图）。对于对话生成任务，系统会根据策略学习生成的系统动作生成相应的回复。

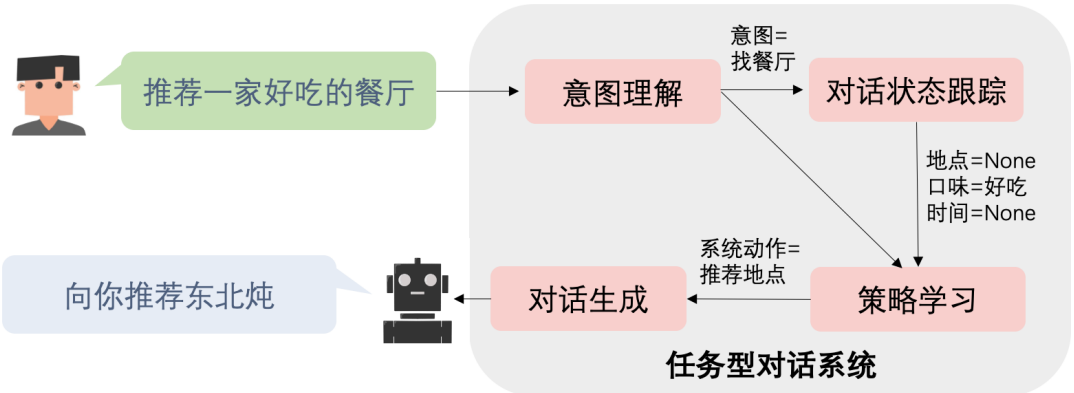

图 2-1　经典任务型对话系统的工作流程。

1　伴随着意图理解的还有槽填充任务，旨在抽取用户输入中的细粒度槽位信息。然而，在多轮对话场景中，对话状态跟踪任务同样可抽取对话过程中的细粒度信息，达到相同目标。因此，本文对槽填充任务不做过多赘述。

2.2 意图理解

意图理解是任务型对话系统的重要组成部分[19-20]，旨在使系统理解用户输入的意图。当系统能够正确识别用户意图时，才能为其提供相应的服务，举例说明，当顾客向手机上的智能助手询问：“推荐一家好吃的餐厅”，智能助手应识别顾客的意图为“找餐厅”。通过结合意图识别的结果以及对对话历史的分析，最终向顾客提供精准回复。当意图理解模块出现错误时，就会出现一步错，步步错的情况，最终导致系统生成与当前上下文不相关的回复，从而影响用户的体验。

意图理解可建模为文本分类任务。随着时代的发展，文本分类的解决方案经历了基于规则、基于统计学习、基于深度学习的时代。

基于规则的文本分类方法是通过人工规定一系列分类规则，例如，根据构建分类词典对文本进行分类[21-24]。这种方法最大的优点是分类过程简单且具有可解释性。然而，这种方法也存在明显的缺陷。首先，建立分类规则需要依赖专家知识，这就需要投入大量的时间和精力来收集和整理专业领域的知识，并将其转化为可用的规则。这个过程对专业领域的经验和知识要求较高，对于非专业人士较为困难，因此建立规则集合需要耗费大量的时间和资源。其次，该方法的泛化性往往较差，这是因为分类规则是人工编写的，所以其泛化性主要限制于规则本身：一旦遇到与规则不匹配的文本，分类效果就会受到严重影响。此外，当规则集合很大时，对规则的更新往往会变得困难。

基于统计学习的文本分类方法是通过对已有的文本数据和标签，来构建分类模型来进行文本分类[25-27]。该方法的主要特点是可以利用大规模的语料库进行训练，从而能够更好地应对复杂的文本分类任务。其中，最具代表性的方法是朴素贝叶斯、支持向量机、最大熵等算法。这类方法都需要进行特征提取，将文本转换为特征向量表示，以便对模型进行学习和分类。在这个过程中，由于特征的选取和提取对分类结果影响很大，需要专家根据具体的任务选择合适的特征，这一过程同样需要耗费时间和资源。

基于深度学习的文本分类方法是利用深度神经网络模型来进行文本分类，是目前最流行的分类方法之一[3,28-29]。相比于基于统计学习的方法，该类方法能够进行端到端的学习，无需进行显式的特征提取，因此在文本分类任务中更高效且准确。常用的深度学习模型包括卷积神经网络、循环神经网络和 Transformer

等。值得注意的是，基于深度学习的方法通常需要大量的标注数据。一些研究工作联合意图分类和其他文本分类任务进行多任务学习，可以有效缓解数据稀疏的问题[30-32]。

以上研究工作都是基于封闭环境下已知意图的分类。在意图理解话题下，还存在一些热门的研究问题：

- **零样本意图分类（zero-shot intent classification）**：该问题旨在提升模型对训练阶段未见过（unseen）意图的分类能力。举例来说，训练阶段的意图包括” 播放音乐”、” 发送短信”、” 查询天气”，而在测试阶段，零样本意图分类关注模型对” 叫出租车” 或” 订机票” 等不在训练集中的意图进行分类。常见的技术路线包括归纳式零样本学习[33]和直推式零样本学习[34]。
- **小样本意图分类（few-shot intent classification）**：该问题的目标是在给定非常有限的意图训练文本（通常为一个或几个）的情况下，使模型能够有效地对这些意图进行分类。例如，在只提供了 5 条属于” 叫出租车” 和” 订机票” 两个意图的用户输入的前提下，训练模型能够对这两个类别进行分类。常见的技术路线包括基于数据增强[35]、元学习[36]和预训练-微调的方法[37]。
- **新意图发现（new intent discovery）**：该问题旨在从大量未标注的用户话语中，通过聚类的方式挖掘出潜在的意图。常见的技术路线包括基于对比学习[38]、半监督学习[39]等。
- **意图持续学习（intent continual learning）**：该问题旨在不重头训练模型的前提下，为模型增加对新意图的分类能力。常见的技术路线包括基于正则[40]、基于重放[41]，以及基于添加新模型结构[42]等方法。
- **分布外检测（out-of-distribution detection）**：这是本文所聚焦的问题，旨在检测出包含训练阶段未曾出现的，未知意图的用户输入。下一章节中将详细论述常见的技术路线。

2.3 分布外检测

传统的机器学习模型往往会基于封闭世界（closed-world）的假设。在该假设下，测试数据和训练数据服从独立同分布，即测试数据都是来自训练分布内的

表 2-1 分布内样本和分布外样本的例子。

编号	文本	标注	分布
1	我看过的最老套的电影	负面	分布内
2	这部电影是一部杰作	正面	分布内
3	给我定个闹钟	未知	语义分布外
4	服务员非常友好	正面	背景分布外
5	这食物咸得不能吃	负面	背景分布外

（in-distribution，简称 ID）。然而，当一个机器学习模型开发完成并上线后，用户输入的内容往往无法得到保证。模型面临着开放世界 (open-world) 的情况，即可能会遇到训练集分布外（out-of-distribution，简称 OOD）的测试数据。因此，评估模型在分布外数据下的表现，提升模型在分布外数据下的稳健性拥有很强的现实意义。

分布外数据可根据语义空间是否发生偏移分为两类：语义分布外（semantic）和背景分布外（background）[43]。语义分布外指的是不属于训练集预定义类别集合的数据分布，例如，对于一个由正面情感（positive）和负面情感（negative）电影影评训练得到情感分类器，一个中性（neutral）的句子就属于语义分布外（表 2-1 中的第三个例子）。背景分布外指的是类别在训练集预定义类别集合内，但该类型数据往往来自不同领域，具有不同的表述风格。举例来说，对于上述的电影影评训练得到的情感分类器，一个餐馆领域的正面或负面的评价数据就为背景分布外（表 2-1 中的第四、五个例子）。这些例子仅针对文本数据，其他模态的数据亦适用。例如，对于一个由生活中的食品图片训练得到的分类器，一张关于人脸的图片就属于语义分布外，一张素描风格的食品图片就属于背景分布外。

语义分布外检测是分布外检测研究工作的焦点。这是由于该类数据已经超出模型的分类能力范畴：其类别不在预定义的类别空间中，模型无法对其分类正确，所以应当对它们提前检测出来，给予预设的操作。对于背景分布外样本，由于其语义空间没有发生偏移，所以模型理论上存在对它们分类正确的能力，主流的工作主要关注模型在它们上的泛化性，即分布外泛化（OOD generalization）[44-47]，不过，亦有工作考察模型对背景分布外检测的能力[43]，这是由于模型尽管理论上存在分类正确的可能，但对于这些样本，模型的表现往往不佳，所以提前检测出它们也可以避免模型犯错。

语义分布外检测（以下简称为“分布外检测”）有众多应用场景。对于任务

型对话系统，分布外检测可以帮助系统识别出用户输入中的未知意图，避免对话系统把它们分类到已知意图中，出现答非所问的情况[48-50]；对于自动驾驶系统，分布外检测可以检测出训练阶段未出现的路标、物体等，并及时把车辆的控制权交由司机处理，避免车辆出现误判和误操作，进而避免事故的发生；此外，分布外检测在科学研究领域也具有广泛的应用前景。例如，在细菌分类模型中，分布外检测可以及时识别出科学界未收录的细菌类型，促进对细菌学的研究和认识[51-52]。

随着深度学习的发展，各种基于神经网络的新颖的分布外检测方法被提出，有工作对这些方法做了较为系统地介绍[53]，其中流行的方法可基于生成式模型[51,54-55], 分类器[56-59]等，这些方法都旨在设计一个打分函数：把输入样本映射为一个分数（score），并给予分布外和分布内样本不同的打分，如对分布内样本打分较高，对分布外样本打分较低。用公式表述如下所示：

$$G(S(\mathbf{x}),\delta)=\begin{cases}\text{ID} & S(\mathbf{x})\geqslant\delta,\\ \text{OOD} & S(\mathbf{x})<\delta,\end{cases} \tag{2-1}$$

这里 G 是分布外检测器，S 是设计的打分函数，$\mathbf{x}$ 是输入样本，可为不同模态，如图片、文本、音频等，δ 是预定义的阈值，用户可根据需求来选取，比如挑选的阈值至少应召回 95% 的分布内数据或分布外样本。

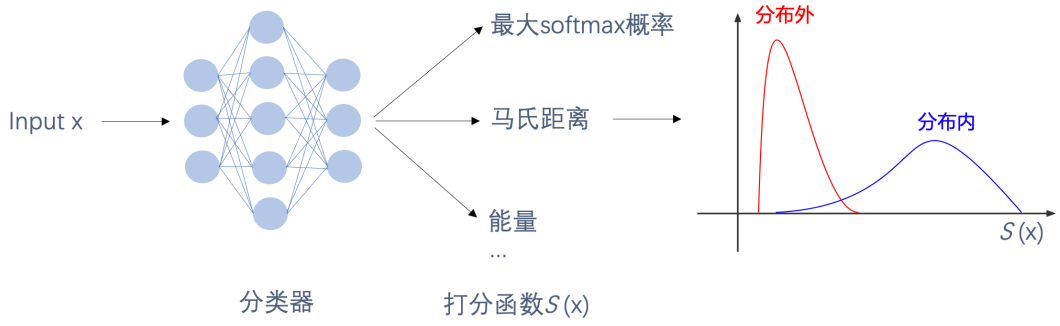

图 2-2 基于分类器的分布外检测

2.3.1 基于分类器的分布外检测

该类方法依赖训练标签的分类器。训练阶段，训练数据是有标注的。分类器结构往往由编码器和全连接层组成，对于文本数据，编码器可以是长短期记忆网络（LSTM）或 BERT 等，对于图像，编码器可以是卷积神经网络（CNN）或 ViT 等[60]。训练目标往往是交叉熵损失。测试阶段，根据分类器的输出来进行分布外检测。

最大 softmax 概率 (maximum softmax probability，简称 MSP)[56]。取测试样本 **x** 在已知标签分布上最大概率值作为打分函数：

$$\mathrm{MSP}(\mathbf{x}) = \max_{y} p(y|\mathbf{x}). \tag{2-2}$$

由于概率值可衡量样本的标签在训练标签集合上的分布，分布内样本的标签属于训练标签集合，所以理想状态下，最大概率值较高，比如接近 1，相反地，分布外样本的标签不属于训练标签集合，所以理想状态下，最大 softmax 概率较低，比如接近 $\frac{1}{N_y}$，这里 N_y 是已知样本标签数量。

MSP 计算简单，而且适用于任何已经训练好的分类器，但后续有工作指出其往往出现过自信（over confident）的问题[61]，即分布外样本的 MSP 值仍较高。该问题一种可能解释是由于分类器在训练阶段，只有分布内样本参与优化，所以其对分布外样本的概率输出无法保证，为解决该问题[61]在训练阶段引入分布外样本进行校准分类器，使得其对分布外样本的输出服从均匀分布，实验证明，校准后的 MSP 可显著提升检测效果。

马氏距离（mahalanobis distance）[57]。取测试样本 **x** 的在特征空间上的表示 $f(\mathbf{x})$（由编码器得到）与训练标签高斯分布的马氏距离作为打分函数：

$$\mathrm{Mah}(\mathbf{x}) = -\max_{y}(f(\mathbf{x}) - \hat{\boldsymbol{\mu}}_y)^T \hat{\boldsymbol{\Sigma}}^{-1}(f(\mathbf{x}) - \hat{\boldsymbol{\mu}}_y). \tag{2-3}$$

这里 $\hat{\boldsymbol{\mu}}_y$ 和 $\hat{\boldsymbol{\Sigma}}$ 是根据训练样本计算得到的经验性均值和协方差矩阵。Lee 等人[57]先是证明了对于由交叉熵训练得到的分类器，样本在其特征空间中的表示都服从协方差矩阵相等的条件高斯分布：不同类别的均值不同，协方差矩阵相同。接着便可以通过样本与训练类别分布的马氏距离来估计样本的概率密度，进

而判断样本是否属于该类别。对于负的马氏距离较小的样本，说明它的概率密度都比较低，即不服从任意类别，所以可判断为分布外样本。

马氏距离不容易受到过自信的影响，比起最大 softmax 概率的性能更好，但后续有工作指出在一些场景下，样本在特征空间上的表示可能不服从高斯分布[62]。虽然马氏距离也适用于任何已经训练好的基于交叉熵的分类器，但 Winkens 等人[63]证明了在训练时引入对比学习，使得相同类别的更近，不同类别的更远，可以进一步提升马氏距离的效果。

能量（energy）[58]。取由分类器推出的测试样本 **x** 的能量值作为打分函数：

$$\text{Energy}(\mathbf{x}) = -T * \log \sum_{y} \exp(f_y(\mathbf{x})/T). \tag{2-4}$$

这里 T 是温度系数，$f_y(\mathbf{x})$ 是指 **x** 样本在 y 标签上的 logit 值（取 softmax 之前的值）。LeCun 等人[64]引入能量来作为负的未归一化的样本的概率密度：样本的能量值越高代表其概率密度越低，Liu 等人[58]证明了基于 softmax 的分类器可看作一个能量模型，计算方式如公式 2-4 所示。样本的能量值越高说明其概率密度也越低，可判断为分布外样本。能量值有理论上的保障，亦不易受到过自信问题的影响，比起最大 softmax 概率的性能更好，成为非常流行的分布外检测方法。为了和其他打分函数保持统一，即分布内分数较高，所以主流工作往往会对能量值取负作为最终的打分函数。

基于分类器的方法虽然都较为简单，但都需要依赖大量的标注数据去训练分类器，这在许多场景下是比较困难且昂贵的，所以有研究者探究在无标注的设置下来检测分布外样本，基于生成式模型的研究路线便是其中一种思路。

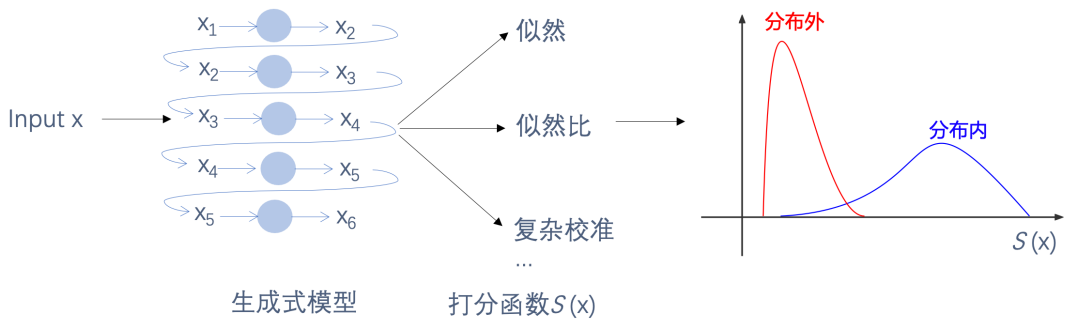

图 2-3 基于生成式模型的分布外检测

2.3.2 基于生成式模型的分布外检测

与分类器不同，生成式模型旨在建模训练数据的分布，在建模方式上，有以GPT 系列为代表的自回归建模[65-66]，也有引入隐变量的变分自编码器[67]，也有生成对抗网络[68]，最近广为关注的 Diffusion 模型[69]。在参数化方面，Transformer 架构占绝对主导[2]。

训练阶段，生成式模型只需要训练样本就可建模其分布，不需要用到样本标签（亦可使用标签来增强建模）。测试时，根据生成式模型的输出来检测分布外样本。

似然（likelihood）[54]。取测试样本 **x** 的似然作为打分函数，以文本为例：

$$\text{likelihood}(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\mathbf{w}_{<i};\theta_0) \tag{2-5}$$

这里 n 表示句中词的个数，x_i 表示 **x** 中的第 i 个词，$\mathbf{w}_{<i} = \{x_0, ..., x_{t-1}\}$，$\theta_0$ 为拟合训练数据分布的生成式模型。

由于生成式模型估计的是训练样本的分布，所以理想状态下，训练集同分布的样本似然值较高，相反地，分布外的样本似然值较低。值得注意的是，该方法在训练和测试时都无需样本标注。

似然方法虽然简单，但后续有工作指出分布内样本的似然高于分布外的假设并不成立，即分布外样本似然值可能仍较高，分布内样本的似然值可能较低[54]。这一违法直觉的现象，启发了后续工作对其的研究。

似然比（likelihood ratio，LLR）[51]。Ren 等人[51]认为 $p(\mathbf{x};\theta_0)$ 失效的原因是其会受到 **x** 中一些与语义无关的背景特征的影响，比如，对于图片数据，背景特征可以是背景像素，对于文本数据，背景特征可以是一些停用词等。为了缓解背景特征的影响，作者使用扰动后的训练样本训练一个生成式模型 θ_1，作者假设扰动后的数据语义特征发生变化，但背景特征没有发生变化，所以 θ_1 可以较好地拟合训练样本中的背景特征，而不去拟合语义特征。在测试时，作者使用两者的比值，来缓解背景特征的影响，进而检测分布外样本：

$$\text{LLR}(\mathbf{x}) = p(\mathbf{x};\theta_0)/p(\mathbf{x};\theta_1). \tag{2-6}$$

似然比方法取得了远超似然的效果，但训练背景模型时也会受到关键超参数的影响：如何扰动以及扰动比例。背景特征的影响不是似然失效的唯一解释，也有工作从输入样本的复杂度展开论证。

复杂校准（complexity calibration）[55]。Serrà 等人[55]通过大量实验证明对于图片数据，仅用 $p(\mathbf{x};\theta_0)$ 会受到图片 **x** 复杂度的影响：复杂度较低的样本似然往往较高，复杂度较高的样本似然往往较低。为了缓解复杂度对似然的影响，测试时作者先通过外部工具衡量图片 **x** 的复杂度 $L(\mathbf{x})$，接着通过比值的方式缓解复杂度的影响：

$$\text{Score}(\mathbf{x}) = p_{\theta_0}(\mathbf{x})/L(\mathbf{x}). \tag{2-7}$$

综上所述，基于生成式模型的分布外检测在得到检测器时往往不需要使用样本的标签，为分布外检测提供另一种可替代的技术路线。但在实验效果上，基于生成式模型的效果往往不如基于分类器的效果，这可能与生成式模型较难训练相关[70]。

分布外检测与以下热门任务有着较为紧密的联系，但不同任务侧重点有所不同[53]：

- **异常检测（anomaly detection，简称 AD）**[71]：该任务旨在检测所有由语义偏移或背景偏移导致的异常点。异常检测假设正常样本（也就是分布内样本）来自同一类别，不关注对正常样本的分类性能。
- **新颖检测（novelty detection，简称 ND）**[72]：该任务与异常检测类似，区别在于（1）重点检测语义偏移导致的分布外样本；（2）分布内样本可以来自多个类别。新颖检测同样不会关注对正常样本的分类性能。
- **开集识别（open set recognition，简称 OSR）**[73]：该任务旨在检测由语义偏移导致的分布外样本，并实现对分布内样本的准确分类，往往聚焦于单标注场景。
- **离群点检测（outlier detection，简称 OD）**[74]。该任务亦旨在检测所有由语义偏移或背景偏移导致的异常点。不同的是，AD、ND 和 OSR 是归纳式的（inductive），即依赖训练和测试范式，训练时，基于训练集来优化模型，测试时，基于训练得到的模型判断异常点。而 OD 是直推式的（deductive），给定所有的样本，直接找出其中的异常点。

- **领域泛化（domain generalization，简称 DG）**[75]。该任务旨在评估模型对由背景偏移导致的分布外样本的分类能力，强调在训练阶段不能接触到测试阶段的分布外数据。
- **领域适应（domain adaptation，简称 DA）**[76]。该任务亦旨在评估模型对由背景偏移导致的分布外样本的分类能力。不同于 DG，该任务在训练阶段可以接触到测试阶段的有标注或无标注的分布外数据。

2.3.3　评估指标

分布外检测本质为二分类任务，即判断用户输入是分布内还是分布外。评价指标往往采用 AUROC、AUPR In、AUPR Out、FPR95[77-79]。

受试者工作特征曲线的面积（area under receiver operating characteristic，简称 AUROC），在计算该指标的过程中，需要先对样本进行排序，并按顺序把样本的预测值当作阈值，划分预测的正负样本，并计算此时的“真正例率”（True Positive Rate，简称 TPR），还有“假正例率”（False Positive Rate，简称 FPR）

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{2-8}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}, \tag{2-9}$$

这里 TP、FN、TN、FP 是根据样本的真实类别与模型预测类别组合的划分。

表 2-2　混淆矩阵

	实际为正例	实际为负例
预测为正例	True Positive (TP)	False Positive (FP)
预测为负例	False Negative (FN)	True Negative (TN)

每次计算完 TPR 和 FPR，就分别以它们为横、纵坐标作图，最终得到“ROC”曲线，AUROC 就是 ROC 曲线的面积，假定 ROC 曲线中各个点的坐标为 $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$，那么 AUROC 可估算为

$$\mathrm{AUROC} = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i)(y_{i+1} + y_i). \tag{2-10}$$

FPR95 是挑选一个阈值使得 TPR 达到 95%，计算此时的 FPR。计算过程中，根据 y 即 TPR 值对 $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ 进行从小到大排序，找到第一个满足 $y > 0.95$ 的点 (x_j, y_j)，此时的 x_j 值就为 FPR95。

精确率-召回率曲线下面积（area under precision-recall，简称 AUPR），在计算该指标的过程中，亦需要先对样本进行排序，并按顺序把样本的预测值当作阈值，划分预测的正负样本，并计算此时的“查准率”（precision，简称 P），还有“查全率”（recall，简称 R）

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{2-11}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{2-12}$$

每次计算完 P 和 R，亦分别以它们为纵、横坐标作图，最终得到“P-R”曲线，AUPR 就是 P-R 曲线的面积。

在 P 和 R 的计算过程，如果把分布内样本当作正样本，最终得到的就是 AUPR In，如果把分布外样本当作正样本，最终得到的就是 AUPR Out。

2.4 本章小结

本章主要针对论文中涉及到的任务型对话系统和分布外检测的背景知识进行介绍。对于任务型对话系统，本章简要介绍了系统的相关模块，详细介绍了其中较为重要的意图理解模块，包括方法的发展历程与研究热点。随后本章介绍了分布外检测的定义、常见的技术路线以及评估指标。

第三章　单标注场景下的分布外检测

3.1　引言

单标注场景下，研究者假设训练和测试阶段中，用户话语仅表达单个意图。例如，话语“今天天气如何?”的意图为“查天气”。基于此假设，如果话语表达的意图是已知的，即属于训练集的意图标签集合，那么该话语就是分布内，相反地，如果意图是未知的，即不属于训练集的意图标签集合，那么该话语就是分布外，是分布外检测的检测目标。

该场景下，训练数据包括用户话语及其对应的意图标签。主流方法利用这些有标注的数据来训练已知意图分类器，并基于该分类器来设计分布外检测器[57-58,61]。具体而言，随着深度学习的发展，这些方法在训练阶段，通常使用交叉熵损失，来训练一个能够对用户话语进行准确分类的神经网络作为分类器。该分类器的输入为用户话语，输出为已知意图。在测试阶段，这些方法会使用基于分类器推导出的话语能量、马氏距离等打分函数来进行分布外检测[57-58]。有关方法的细节，可查看第 2.3 节。

尽管以上方法能取得不俗的性能，本工作发现，当仅使用已知意图的训练数据来训练分类器，并作为分布外检测器时，它会受到词级别偏差特征 (word-level bias feature) 的过度影响。也就是说，分布外检测器往往会根据句中的词级别的特征，而非句子的意图特征，来决定句子是否为分布外。举例说明，分布外检测器只要识别到句中包含“text”一词，就把它判定为表达已知意图“发短信”的分布内样本，尽管句子的意图是未知的，与“发短信”毫无关系。

分布外检测器受到词级别偏差的原因之一是训练集中，训练样本中的某些词和其意图标签具有强相关性（strong correlation）。例如，在 CLINC150 的训练集中[80]，所有带有“text”一词的句子都被标注为“发短信”（如表 3-1 中的例子 1，2 所示）。基于这种强相关性，分类器在训练过程中，仅仅通过样本中的“text”这一简单的词级别特征就能够正确分类句子的意图，而且不会影响验证和测试

表 3-1　带有“text”词的分布内样本和分布外样本。

序号	用户话语	意图	来源
1	text wenona and tell her we will be there tomorrow	发短信	训练集
2	i have to send a text	发短信	训练集
3	send davis a text i’ll be there in a few	发短信	测试集
4	please read the text message i just received	未知（分布外）	测试集
5	save my text on my laptop hard drive	未知（分布外）	测试集

集上，对“发短信”类别句子的分类性能，因为这些句子与训练集存在着独立同分布的假设，即带有“text”词的句子的意图标注也都是“发短信”（表 3-1 中的例子 3）。

需要注意的是，训练样本中的词和意图标签的强相关性往往是虚假的（spurious），并不具备可泛化性，即不是在任何数据分布下都成立。这是因为意图分类是一个句子级别的任务，只有句子的整体语义才能决定其意图，而词级别特征无法单独决定句子的意图。以前述例子为例，单独的“text”一词无法确定意图为“发短信”，只有结合上下文才能共同决定。如果分类器仅仅基于“text”这一偏差特征做决策，那么由该分类器推导出的检测器就会把带有“text”的分布外样本误分类成分布内样本（表 3-1 中的例子 4，5）。举例来说，基于马氏距离的检测器是根据分类器得到的句子表示进行分布外检测，如果分类器仅仅关注“text”一词，那么带有“text”特征的分布外和分布内句子在空间上的特征表示就会相近，无法得到有效区分。

为了缓解词级别偏差（简称“偏差词”）对分布外检测的影响，本文先提出一种偏差词识别方案，接着依据识别到的偏差词来检索分布外文本，最后基于检索到的文本校准检测器，达到缓解偏差的目标。具体而言，本文基于梯度的方法来识别对检测器影响较大的词作为偏差词，接着从公开的分布外数据集中检索带有相同偏差词的文本，最后根据检索的分布外文本，设计新的训练目标来校准检测器，使得检测器对包含相同偏差词的分布外和分布内样本打分不同，以此来缓解偏差的影响。

本文的贡献可以总结如下：

- 本文首先论证了在单标注场景下，分布外检测器容易受到词级别偏差的影响，进而误将分布外样本分类为分布内。为解决这一问题，本文提出了一种基于分布外数据增广的校准方案，旨在缓解偏差词的影响。

- 实验方面，本文在不同的分布外检测器和数据集上进行了验证，经过本文提出的方法进行校准后，所有评估指标都得到了稳定的提升。此外，校准后的模型对于分布内样本的已知意图分类性能并未受到影响。

3.2 相关工作

分类任务中的偏差特征。在深度学习时代，交叉熵损失已成为许多自然语言分类任务中模型的常规优化目标。然而，一些研究者发现，仅通过优化交叉熵损失得到的分类器可能会根据一些偏差特征（bias feature）做决策，而非根据决定类别的因果特征[81-85]。例如，在自然语言推理任务中（Natural Language Inference，简称 NLI，旨在分类两个输入句子的语义关系），Gururangan 等人[81]对 SNLI 和 MNLI 数据集[82]进行了统计分析，他们发现一些简单特征，例如句中是否包含否定词，与“矛盾”类别高度相关。同时，模型学习到了这些简单特征，并直接利用它们来判断两个句子之间的关系，而不是基于句子的语义特征。由于这些简单特征并不是我们期望模型学习和利用的特征，因此被称为偏差特征。类似的现象也在情感分类、事实核查和释义识别等任务中被其他研究人员发现[83-85]。偏差特征是决定句子标签的非因果特征，不具备泛化性。在 NLI 任务中，否定词就是一个例子，它不能单独决定两个句子之间的关系。尽管依赖于偏差特征的分类器在与训练集同分布的样本上表现良好（因为训练集中存在这种偏差），但在不包含该偏差的数据分布下，模型的分类性能会下降[81,83]。例如，通过在 MNLI 数据集上训练的 NLI 模型，在 HANS 数据集上的分类性能通常较差（因为否定词不再决定矛盾类别）。与之前的研究不同，本文首次探究了偏差特征对分布外检测任务的影响，并提出了一种基于分布外数据校准的方法来缓解偏差的影响。

分布外检测器校准。已经有工作在训练过程中，通过引入分布外样本，来增强分布外检测器的性能[58,61,86]，与这些工作不同的是，本文详细阐述了引入分布外样本参与训练的原因：缓解偏差特征的影响，并且基于这一原因，设计了新的分布外样本的检索和校准方法。

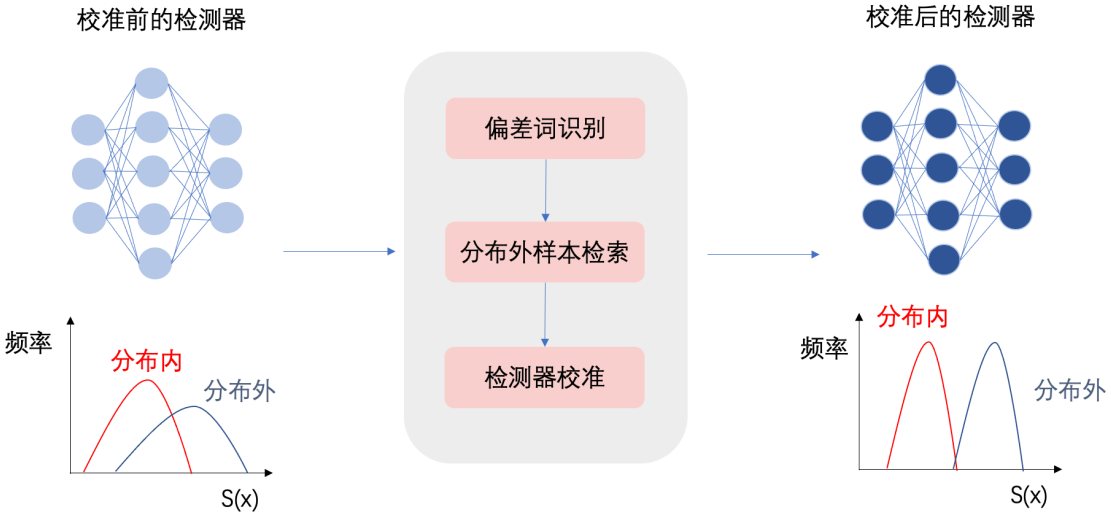

图 3-1　校准流程示意图。

3.3　本文方法

为缓解偏差词对分布外检测器的影响，本章节中，我们先是对检测器依赖的偏差词进行识别（第 3.3.1 小节），之后，基于识别的偏差词来检索公开的分布外样本（第 3.3.2 小节），最终利用检索得到的分布外样本，设计新的训练目标对检测器进行校准（第 3.3.3 小节）。整体示意图如图 3-1 所示。

3.3.1　偏差词识别

本文设计新的偏差函数 $I(w,S)$，来为检测器 S 识别出偏差词 w：$I(w,S)$ 越高，对于 S，w 越有可能成为偏差词。$I(w,S)$ 的计算依赖于打分函数 S 对词 w 向量的梯度，所以只要打分函数 S 的梯度可以反向传播至词向量，其计算细节不作要求，可以是能量、马氏距离等。具体而言，给定训练集中的每个句子 $\mathbf{x}$, 通过 $S(\mathbf{x})$ 对句中词 w 的词向量 $\mathbf{e}_w$ 的梯度与自身的乘积，作为词 w 在句子 $\mathbf{x}$ 中偏差程度的分数 $I(w,S,\mathbf{x})$：

$$\begin{aligned} I(w,S,\mathbf{x}) &= \mathbf{e}_w \nabla \mathbf{e}_w \\ &= \mathbf{e}_w \frac{\partial S(\mathbf{x})}{\partial \mathbf{e}_w}, \end{aligned} \tag{3-1}$$

$I(w,S,\mathbf{x})$ 可近似估计把句子 $\mathbf{x}$ 中词 w 的词向量 $\mathbf{e}_w$ 置为 $\mathbf{0}$ 后，检测器 S 对句子打分的变化[87]，越高表示词 w 对 $S(\mathbf{x})$ 的提升越大，即 S 在判断 $\mathbf{x}$ 是否为分布

Algorithm 1 偏差词识别

Input: 已知意图的训练集合 $\mathcal{X}_{train}^{in}$，阈值 θ，打分函数 S。

```
I = {}  # 记录对于 S，词 w 的偏差程度 I(w, S)
for x ∈ X_train^in do
    for w ∈ x do
        I[w] = 0  # 初始化
    end for
end for
for x ∈ X_train^in do
    计算 S(x)
    for w ∈ x do
        根据公式 3-1 计算 I(w, S, x)
        I[w] = I[w] + I(w, S, x)
    end for
end for
B = {}  # 记录偏差词
for w ∈ I do
    if I[w] > θ then
        Add w into B
    end if
end for
Return B
```

外样本时对 w 非常依赖。

本文对原有训练集 $\mathcal{X}_{train}^{in}$ 中所有词，计算它们对所在句子得分的影响，汇总后作为词 w 对 S 的综合影响 $I(w, S)$：

$$I(w, S) = \sum_{\mathbf{x} \in \mathcal{X}_{train}^{in}} \sum_{w_i \in \mathbf{x} \wedge w_i = w} I(w, S, \mathbf{x}) \tag{3-2}$$

这里采用求和方式进行汇总是融入词 w 出现次数的影响。本文也实验了求平均的方式，实验结果证明求和的方式效果更好，这是由于求和也会受词 w 出现次数的影响，而求平均则消除了次数的影响。最终本文选择 $I(w, S)$ 较高的词作为偏差词（高于给定阈值），最终得到偏差词集合 **B**，整个识别过程如Algorithm 1 所示。

3.3.2 分布外样本检索

得到偏差词集合 **B** 后，本文接着根据这些偏差词，来针对性地收集相关的分布外样本校准检测器。具体而言，本文从公开的包含大量分布外数据的集合 **D**

Algorithm 2 分布外样本检索算法

Input: 外部收集的分布外数据集 **D**，偏差词集合 **B**，每个偏差词检索的分布外数目 k。

$\mathcal{X}_{train}^{out} = \{\}$ # 记录检索的样本
Count $= \{\}$ # 记录每个偏差词检索的个数
for $w \in$ **B do**
    Count$[w] = 0$ # 初始化
end for
for $\hat{\mathbf{x}} \in$ **D do**
    for $w \in \hat{\mathbf{x}}$ **do**
        if $w \in$ **B** and **Count**$[w] < k$ **then**
            Add $\hat{\mathbf{x}}$ into $\mathcal{X}_{train}^{out}$
            Count$[w] =$ **Count**$[w] + 1$
        end if
    end for
end for
Return $\mathcal{X}_{train}^{out}$

中，检索一定量的包含这些偏差词的分布外样本组成集合 $\mathcal{X}_{train}^{out}$ 来校准检测器，具体检索过程如Algorithm 2 所示。

3.3.3　检测器校准

检测器校准旨在引入新的训练目标，缓解集合 **B** 中的偏差词对检测器的影响。新的训练目标的主要思想是使得检测器对包含相同偏差词分布外样本 $\mathcal{X}_{train}^{out}$ 和分布内样本 $\mathcal{X}_{train}^{in}$ 的打分有间隔，如此一来，检测器就不会受到偏差词的过度影响，这是因为偏差词在分布外和分布内样本都存在，为了使得这些样本打分尽可能不同，检测器不能只关注它们间的共有特征，即偏差词，还要关注它们间的不同特征，即意图特征。具体而言，新的训练目标如下所示：

$$\begin{aligned}\mathcal{L}_{\text{aux}} =& \mathbb{E}_{\mathbf{x}\sim\mathcal{X}_{train}^{in}} \max(0, m_{\text{in}} - S(\mathbf{x})) \\ &+ \mathbb{E}_{\hat{\mathbf{x}}\sim\mathcal{X}_{train}^{out}} \max(0, S(\hat{\mathbf{x}}) - m_{\text{out}}),\end{aligned} \tag{3-3}$$

这里通过 hinge loss 使得打分有间隔：分布内的样本打分函数都高于阈值 m_{in}，分布外样本的打分函数都低于阈值 m_{out}。

校准后，新的打分函数 S 可以直接用作分布外检测，细节可参考公式 2-1 。

表 3-2 数据集 CLINC150 和 CLINC150-IMB 的统计信息。

Statistic	CLINC150	CLINC150-IMB
Train-ID	15,000	10,525
Validation-ID	3,000	3,000
Validation-OOD	100	100
Test-ID	4,500	4,500
Test-OOD	1,000	1,000
Number of known intents	150	150

再次强调，校准的整个过程可适用于任何可微的打分函数 S，本文的实验中针对马氏距离和能量两个打分函数进行测试，以展现校准方法的通用性。

3.4 实验设置

3.4.1 数据集

为了评估本文提出的校准方法在单标注分布外检测上的表现，本文在该场景下的两个标准数据集 CLINC150 和 CLINC150-IMB[80]进行评测：

- CLINC150 是一个涵盖了语音助手场景下可能出现的 150 种意图的数据集。对于每个意图，训练集中提供了 100 个用户话语作为样本。其中分布外数据是具有未知意图的用户话语。
- CLINC150-IMB 是 CLINC150 的类别不均衡版本，可以进一步评测我们的方法在类别不均衡数据上的表现。该数据集的训练集中每个意图的用户话语数目不一定为 100，可为 25，50，75，或 100，但验证集和测试集与 CLINC150 相同。

表 3-2 提供了两个数据集的统计数据。Train-ID 是指训练集中的分布内句子，其他类似，Number of known intents 是分类器的类别数目。

3.4.2 基线方法

本文使用了三个非常有竞争力的方法，包括最大 softmax 概率（MSP），能量（Energy），马氏距离（Mahalanobis distance）作为基线方法，这三种方法都可以由基于交叉熵训练得到的分类器直接推出，更多关于细节可查看第 2.3 节。

- MSP 取测试样本在已知意图分布上最大 softmax 概率值作为打分函数。

- Energy 取测试样本的能量值作为打分函数。
- Mahalanobis distance 取测试样本的在特征空间上的表示与训练标签高斯分布的马氏距离作为打分函数。

3.4.3　实现细节

为了保证公平对比，所有基线方法同本文的校准方法都基于预训练的 BERT[88]，句子向量都采用 [CLS] 的隐层表示，本文根据验证集的 AUROC 指标来挑选超参数和进行早停策略。本文使用了 5 个种子 $\{0, 1, 2, 3, 4\}$ 对所有结果进行平均。

关于本文提出校准方法的超参数。在词级别偏差识别时，超参是偏差词的数目，即 **B** 集合的大小，本文取 $I(w, S)$ 较高的前 $\{100, 200, 300, 400, 500, 600\}$ 个词分别组成 **B**。在分布外样本筛选时，超参是为每个偏差词筛选的分布外数目 k，本文取 k 时，经验性地保证 k 和 $|\mathbf{B}|$ 的乘积等于训练集分布内数目，即使得参与训练的分布内和分布外样本数目相等。关于检测器校准，对于马氏距离 $m_{\mathbf{in}}$ 为 0，即使得分布内样本尽可能靠近其类别中心，对于 $m_{\mathbf{out}}$，本文提出了一种启发式方法，对分布内样本的距离从小到大进行排序，取 95% 位置的距离作为阈值，这么做的目的是使得 95% 的分布内样本的距离都要小于分布外样本，当然，这里也可以为 $m_{\mathbf{out}}$ 直接预设超参。对于能量，本文受 Liu 等人[58]的启发，直接把 $m_{\mathbf{in}}$、$m_{\mathbf{out}}$ 当作超参进行搜索。本文受 Chen 等人[89]的启发，采用 MIX 数据集来作为分布外数据集。

3.5　结果和分析

3.5.1　主结果分析

主实验结果如表 3-3，表 3-4，表 3-5 所示，所有的结果都是百分比。加粗的结果是最好的结果，底部加横线的结果是次好的结果。由此可以观察到：

- 本文提出的校准方法可以在不同数据集，不同打分函数下提升分布外检测的表现。具体而言，在两个数据集上，使用本文提出的方法对基于马氏距离和能量的检测器校准后，所有分布外检测的指标都得到了稳定提升，表 3-5 亦展示了不同句子表示下方法的有效性，可进一步体现方法的通用性。

- 相较于能量，本文提出的校准方法对于马氏距离的提升更大。由两表可知，校准前，能量比起马氏距离有明显的优势：两个数据的所有指标的比较中，能量的表现都要更好，在校准后，马氏距离的劣势得以缩小，甚至在 CLINC150-IMB 数据集上，两者的性能各有强弱，这是由于马氏距离更容易受到词级别偏差的影响，本文将在第 3.5.2 小节中详细论证。

表 3-3 CLINC150 数据集上不同方法的表现。

Methods	CLINC150			
	AUROC ↑	**FPR95↓**	**AUPR In↑**	**AUPR Out↑**
MSP	96.21 ± 0.22	15.95 ± 2.86	99.11 ± 0.07	85.35 ± 0.75
Energy	97.89 ± 0.02	8.92 ± 0.27	99.48 ± 0.02	92.36 ± 0.14
Mahalanobis	97.46 ± 0.05	10.23 ± 0.41	99.40 ± 0.02	89.94 ± 0.14
Energy + Calibration (ours)	**98.48 ± 0.03**	**5.38 ± 0.23**	**99.60 ± 0.01**	**94.69 ± 0.11**
Mahalanobis + Calibration (ours)	98.33 ± 0.02	5.89 ± 0.24	99.57 ± 0.01	94.12 ± 0.16

表 3-4 CLINC150-IMB 数据集上不同方法的表现。

Methods	CLINC150-IMB			
	AUROC ↑	**FPR95↓**	**AUPR In↑**	**AUPR Out↑**
MSP	95.52 ± 0.17	19.04 ± 1.49	98.93 ± 0.05	83.51 ± 0.32
Energy	97.55 ± 0.07	10.82 ± 0.26	99.40 ± 0.02	91.21 ± 0.24
Mahalanobis	97.05 ± 0.11	12.01 ± 0.52	99.30 ± 0.02	88.60 ± 0.42
Energy + Calibration (ours)	98.09 ± 0.13	**8.31 ± 1.16**	99.54 ± 0.03	**92.81 ± 0.60**
Mahalanobis + Calibration (ours)	**98.13 ± 0.07**	8.4 ± 0.47	**99.58 ± 0.01**	92.40 ± 0.54

表 3-5 CLINC150 数据集上不同方法的表现，句子表示由平均池化得到。

Methods	CLINC150			
	AUROC ↑	**FPR95↓**	**AUPR In↑**	**AUPR Out↑**
MSP	96.0 ± 0.2	17.3 ± 2.6	99.0 ± 0.1	85.1 ± 0.7
Energy	97.6 ± 0.0	10.2 ± 0.4	99.4 ± 0.0	92.0 ± 0.3
Mahalanobis	97.4 ± 0.1	10.5 ± 0.6	99.4 ± 0.0	89.6 ± 0.6
Energy + Calibration (ours)	**98.2 ± 0.2**	**7.5 ± 0.8**	**99.6 ± 0.1**	**94.0 ± 0.3**
Mahalanobis + Calibration (ours)	98.1 ± 0.0	7.6 ± 0.5	99.5 ± 0.0	93.3 ± 0.3

3.5.2 偏差函数分析

为了论证偏差词对分布外检测器的影响，和偏差函数 $I(w, S)$ 对识别偏差词的有效性，本文先通过 $I(w, S)$ 识别出测试集中每个分布外句子的最有可能的偏

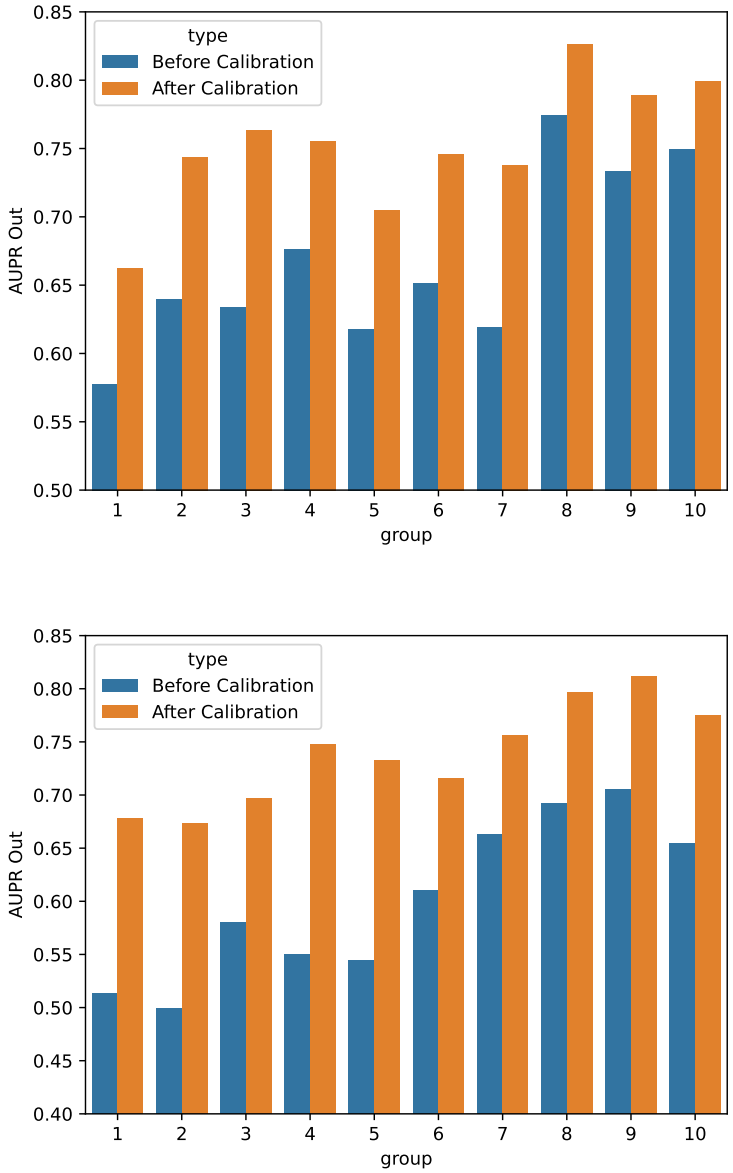

图 3-2　CLINC150 数据集下，不同组的能量（上图）和马氏距离（下图）的性能。

差词，并基于该词的 $I(w,S)$ 对分布外句子进行从高到低排序，分成十组：第一组句子的最大 $I(w,S)$ 最高，即偏差词特征最明显，第十组句子的最大 $I(w,S)$ 最低，即偏差词特征最不明显，最后分别评估检测器对这些组的分布外样本的检测性能。

结果如图 3-2 所示，可以看到校准前检测器对句中包含 $I(w,S)$ 较高的分布外句子较难检测：第一组（gropu 为 1）的检测性能普遍较低，随着组数增加（group 升高），检测性能呈上升趋势。这说明 $I(w,S)$ 高的词，更容易成为偏差词，导致模型的性能下降。而校准后，所有组的性能都得到了提升。

表 3-6 基于偏差词采样和随机采样的分布外数据集在 CLINC150 上的性能对比。

Methods	CLINC150			
	AUROC ↑	**FPR95↓**	**AUPR In↑**	**AUPR Out↑**
Energy + Random	97.96 ± 0.08	7.99 ± 0.59	99.48 ± 0.03	92.83 ± 0.12
Energy + Calibration (ours)	**98.48 ± 0.03**	**5.38 ± 0.23**	**99.60 ± 0.01**	**94.69 ± 0.11**
Mahalanobis + Random	98.06 ± 0.05	7.21 ± 0.37	99.50 ± 0.01	93.41 ± 0.18
Mahalanobis + Calibration (ours)	**98.33 ± 0.02**	**5.89 ± 0.24**	**99.57 ± 0.01**	**94.12 ± 0.16**

表 3-7 基于偏差词采样和随机采样的分布外数据集在 CLINC150-IMB 上的性能对比。

Methods	CLINC150-IMB			
	AUROC ↑	**FPR95↓**	**AUPR In↑**	**AUPR Out↑**
Energy + Random	97.93 ± 0.06	9.20 ± 0.67	99.48 ± 0.01	92.67 ± 0.25
Energy + Calibration (ours)	**98.09 ± 0.13**	**8.31 ± 1.16**	**99.54 ± 0.03**	**92.81 ± 0.60**
Mahalanobis + Random	97.74 ± 0.09	9.66 ± 0.28	99.43 ± 0.02	92.20 ± 0.30
Mahalanobis + Calibration (ours)	**98.13 ± 0.07**	**8.4 ± 0.47**	**99.58 ± 0.01**	**92.40 ± 0.54**

同时，我们也可以看到，比起能量，马氏距离的前面组的分布外性能较低，比如能量在 CLINC150 数据集第一组的 AUPR Out 为 57.75%，而马氏距离为 51.31%，说明马氏距离更容易受到词级别偏差的影响，而校准后两者能获得接近的性能 (66.27% vs. 67.80%)。

为了进一步证明 $I(w, S)$ 的有效性，本文在筛选分布外样本时，不再依据 $I(w, S)$，而是随机采样相同数量的分布外样本对检测器进行校准，结果如表 3-6 和表 3-7 所示，可以看到，本文的方法在不同数据集、不同打分函数、不同指标上都要好于随机采样的分布外样本，这说明由 $I(w, S)$ 提取的分布外样本，对提升模型的检测性能更具有针对性。

3.5.3 分布内类别分类分析

校准后的模型不会影响对已知意图的分类能力。除了分布外检测性能外，本文也测试了校准后的模型参数是否会影响其对分布内用户话语的已知意图分类能力，表 3-8 展示了校准前后两个数据集上的测试集准确率，可以观察到，模型的已知意图分类能力并未受到明显影响，即未出现明显的性能下降，甚至在 CLINC150-IMB 数据集上的分类性能也略有提升，这说明校准后的模型可以同时作为已知意图分类器和分布外用户话语检测器。

表 3-8　分布内用户话语分类准确率

Methods	CLINC150	CLINC150-IMB
Energy	96.38 ± 0.15	95.58 ± 0.17
Energy + Calibration	96.34 ± 0.13	96.36 ± 0.07
Mahalanobis	96.07 ± 0.14	95.25 ± 0.18
Mahalanobis + Calibration	95.84 ± 0.10	95.64 ± 0.16

表 3-9　带有“text”词的样例分析。

序号	方法	用户话语	能量	阈值	预测
1	校准前	send davis a text i’ll be there in a few	18.79	12.82	分布内 ✔
1	校准后	send davis a text i’ll be there in a few	23.60	18.00	分布内 ✔
2	校准前	save my text on my laptop hard drive	12.95	12.82	分布内 ✖
2	校准后	save my text on my laptop hard drive	15.93	18.00	分布外 ✔
3	校准前	please read the text message i just received	13.74	12.82	分布内 ✖
3	校准后	please read the text message i just received	17.85	18.00	分布外 ✔

3.5.4　样例分析

为了直观展现本文提出的校准方法的有效性，本文在表 3-9 中展示了 CLINC150 数据集上的带有偏差词“text”的样例分析。

第一个样例表示对于带有“text”词的分布内样本，校准前后的检测器都有能力正确识别，即输出的能量值高于阈值（这里取了能召回 95% 的分布内样本的能量作为了阈值）。比较校准前后在第二个和第三个样本可以发现，校准前的模型都把这样的分布外的样本误分类成了分布内，即输出的能量亦高于阈值。而校准后的检测器则进行了正确预测，即输出的能量小于阈值。

3.5.5　超参分析

本文提出的方法超参对偏差词数目不敏感。偏差词数目，即集合 **B** 的大小是本文提出的校准方法的关键超参，为了验证它对校准性能的影响，本文在图 3-3 展示了不同偏差词数目下（横坐标），分布外检测的性能（纵坐标），校准结果的变化由蓝色实线绘制，红线为校准前的性能。由图中可以看到，蓝色实线在不同设置下的各个偏差词数目的性能表现都要好稳定好于校准前的性能，证明了本文提出方法对超参数的稳健性。

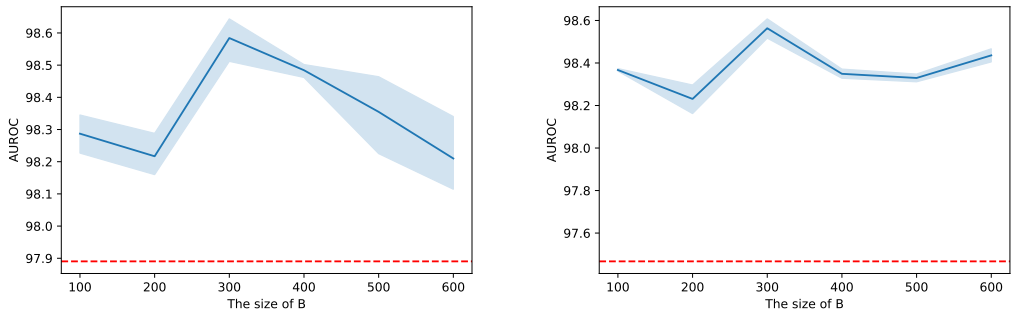

图 3-3 CLINC150 数据集下能量（左）和马氏距离（右）的性能。

3.6 本章小结

本文论述了单标注场景下的分布外检测器，往往会受到词级别偏差的影响，进而影响检测性能。为了解决该问题，本文提出了一种基于梯度的词级别偏差识别方式，接着，基于识别得到的偏差词，本文从外部的公开分布外数据集中检索相关的样本，最后提出了新的损失函数来校准检测器。实验上，本文在不同打分函数、数据集、评估指标上验证了校准的有效性，全面的分析也展现了校准方法的稳健性，即不易受到超参的影响，而且不会影响对已知意图的分类性能。

尽管本文提出的方法能够有效识别和校准文本中的词级别偏差，但它仍存在一定的改进空间：

- 除了词级别的偏差，其他级别的偏差，例如短语和语法级别等，可能也会分布外检测产生影响。因此，在未来的研究中，研究人员可以全面探索和消除不同级别的偏差的影响。
- 不借助分布外样本的情况下消除词级别偏差的影响。仅利用分布内样本来消除偏差是一个值得探索的方向，这样就可以减轻对分布外样本的收集开销。

本文的部分内容已总结成文：

- **Yawen Ouyang**, Jiasheng Ye, Yu Chen, Xinyu Dai, Shujian Huang, and Jiajun Chen. “Energy-based Unknown Intent Detection with Data Manipulation,” in *Proceedings of the 59th annual meeting of the association for computational linguistics.* (**Findings of ACL 2021**)

但与已发表的版本相比，本文做了以下改进：

- 引入更多的打分函数：已发表版本中仅在能量上进行了实验，而本章节的工作在能量和马氏距离上都进行了实验，进一步展示了提出方法的通用性。
- 简化了分布外样本增广方案：已发表版本中的获取方案需要使用外部预训练模型，整个过程相对繁琐。本章节的工作对该方案进行了改进，去除了外部模型的需求，使得整个过程更加简洁。
- 更高的性能：已发表工作在 CLINC150 数据集上的 AUROC 性能仅达到 97.3%，而新的校准方法在 AUROC 上的性能提升至 98.2%。

第四章　多标注场景下的分布外检测

4.1　引言

以往关于分布外检测的工作往往聚焦于单标注场景，即假设每个用户输入只表达单个意图。基于此假设，流行的策略是使用分类器的输出来判断用户输入是否包含未知意图。根据输出是否超过预定义的阈值，可以将输入分类为分布内或分布外。例如，常用的策略包括使用最大 softmax 概率或马氏距离等指标来实现这一目的。

尽管单标注场景非常流行，但在实际应用中，这种假设往往不能得到保证。在真实场景中，用户输入通常会表达一个或多个意图。事实上，Gangadharaiah 等人[90]对亚马逊内部意图数据集的研究后指出，52% 的用户话语具有多个标注，这显示了多标注场景更符合实际情况。然而，尽管多标注场景更贴近现实，但在分布外检测任务上却一直没有被充分地研究和探索。

尽管多标注场景尚未得到充分探索，我们可能对以下问题感到好奇：在多标注场景下，相比于单标注场景，存在哪些新的挑战？我们是否能够直接将单标注场景下流行的方法（如基于分类器的方法）迁移到多标注场景中？下文对这些问题进行了初步探索。

本文提出了多标注场景下独有的挑战，即用户话语可能同时表达已知和未知意图，因为该场景下的意图之间并不互相排斥。我们称这类用户话语称为混合分布外（mixed out-of-distribution，简称 mixed OOD），举例说明，对于疫情前的导航领域的语音助手，混合分布外的例子可以是“导航去南京，帮我向那里的社区做个报备”、“导航去南京，沿路帮我预约个核酸”，在这些例子中，“导航”是已知意图，而“做报备”、“预约核酸”则是未知意图。相反地，对于那些只包含未知意图的用户话语，我们称之为纯净分布外（pure out-of-distribution，检测 pure OOD），比如“给社区做个报备，顺便预约个核酸”。值得注意的是，在单标注场景下，所有包含分布外用户意图的用户话语都属于纯净分布外，因为

它们只能包含单个意图。我们只需要检测出纯净分布外的用户话语即可。然而，在多标注场景，我们需要同时检测出纯净分布外和混合分布外的用户话语。

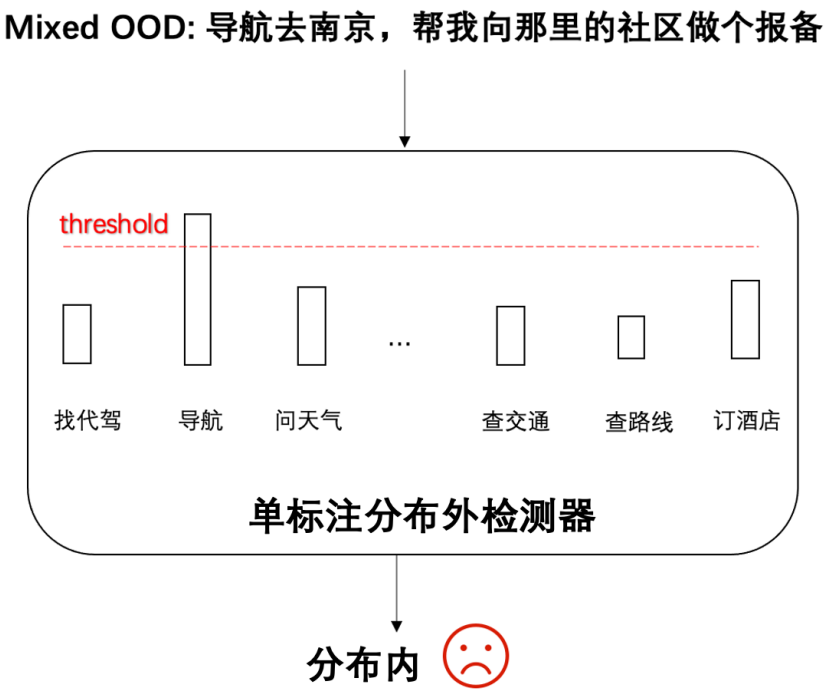

图 4-1　混合分布外被单标注检测器误分类成分布内。

由于混合分布外样本这一新挑战的存在，为单标注设计的检测器不能直接在多标注场景下使用。混合分布外样本中由于存在已知意图，单标注检测器会检测到该意图，并产生大于阈值的输出，导致错误预测，如图 4-1 所示，检测器在检测到“导航”这一已知意图后，就会把该句误分类成分布内，从而忽略了句中的其他未知意图，如“做报备”。另外，单标注场景下的检测器往往基于 softmax 交叉熵进行训练，这种损失函数要求样本中只能包含单个标签，因此，在多标注场景下，单标注场景下的检测器的训练也面临困难。

为解决以上挑战，本文提出了一个基于意图数目协助的新方法 AIK(whether **A**ll **I**ntents contained in the utterance are **K**nown) 来同时检测混合和纯净分布外的用户话语。AIK 的主要思路是首先预测用户话语中的意图数目，接着来检查用户话语中是否包含相同个数的已知意图。如果用户话语中不包含相同个数的已知意图，即说明话语中包含未知意图，该话语即可被分类成分布外。技术细节上，AIK 假设已知意图相关的表示服从条件高斯分布，那么 AIK 就可以提取用户话语中的已知意图相关的表示，接着衡量其与相应的条件高斯分布的马氏距离来判断用户话语是否包含该已知意图。实验上，本文证明了 AIK 可以显著提

Algorithm 3 基于 AIK 的多标注分布外检测

Input: 测试句子 **u**; 阈值 τ; 已知意图集合 Y_{in}，每个已知意图 c 的分布 $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$.

预测句子 **u** 的意图数目 r
D = {} # 记录负的马氏距离
for $c \in Y_{\text{in}}$ **do**
    提取句子 **u** 的和已知意图 c 相关的句子表示 $\mathbf{v}_c$
    计算 $\mathbf{v}_c$ 和 $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$ 的马氏距离 d_c
    加入 $-d_c$ 到 **D**
end for
$S(\mathbf{u}) = r$-th maximum **D**
if $S(\mathbf{u}) < \tau$ **then**
    return 分布外
else
    return 分布内
end if

高对分布外样本的检测能力，尤其是对于混合分布外样本。该工作的代码已经开源在https://github.com/yawenouyang/AIK。

本文的贡献总结如下：

- 本文将对话场景下的分布外检测，从单标注场景扩展到了多标注场景，使该任务更贴合真实场景下的任务型对话系统。同时，本文也提出了多标注场景下的新挑战，即对混合分布外用户输入的检测。
- 本文为多标注场景提出了一个新颖并有效的方法 AIK，通过识别句中的意图是否都是已知的，AIK 可以自然地检测出纯净和混合的分布外用户输入。
- 本文展示了 AIK 在多标注数据集上的比现有方法的优越性。比如，在 MultiWOZ2.3 数据集上，相较于现有方法，AIK 可以将 FPR95 指标显著减低 12.25%。

4.2 本文方法

多标注分类旨在为每个样本分配多个非互斥的标注，针对文本模态的数据，已经有许多杰出的方法被提出来解决这个问题[91-93]，如二元关联（Binary relevance）[94]，该方法将每个标签看作一个单独的二分类问题，并为每个标签训练一个二分类器；分类器链（Classifier chains）[95]，该方法与二元关联类似，不同之处在于，分类器链会对标签进行排序，后面的标签预测会考虑到前面标签的输出，以考虑标签之间的相关性；序列到序列（sequence-to-sequence，简称 seq2seq）

模型[96]，该思路把多标注分类建模成文本生成问题，通过训练一个序列生成模型，输入文本，输出为文本的标签序列。多标注意图分类也最近引起了人们的关注[97-98]。然而，这些方法都基于封闭世界的假设，即假设测试样本的所有类别都是已知的。

在现实世界中，未知标签的出现是不可避免的，不同于以往的工作，本文基于开放世界的假设，并检测带有未知标签的输入，为此，本文提出了一个新颖的方法 AIK。在本章中，本文首先形式化任务，接着介绍 AIK 的整体思想，最后展示它的模型结构和训练目标。

4.2.1　任务形式化

多标注分布外检测打破了用户话语中只能包含一个意图的假设，允许一个用户话语包含一个或多个意图。它旨在检测出包含未知意图的分布外用户话语。

形式化上，给定训练集 $\mathcal{D} = \{(\mathbf{u}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$，这里 $\mathbf{u}^{(i)}$ 是用户话语, $\mathbf{y}^{(i)}$ 是 $\mathbf{u}^{(i)}$ 中表述的意图集合，它包含于已知意图集合 $Y_{\text{in}} = \{c_1, c_2, ..., c_k\}$ 中，也就是 $\mathbf{y}^{(i)} \subseteq Y_{\text{in}}$。测试时，给定一个用户话语，如果它的意图集合 $\mathbf{y}$ 中所有意图不都属于已知意图集合 Y_{in} 中，即它包含未知意图，那么它就是分布外。更进一步，如果 $\mathbf{y} \cap Y_{\text{in}} \neq \varnothing$，说明该分布外用户话语同时包括已知意图，那么它被称为混合分布外，如果 $\mathbf{y} \cap Y_{\text{in}} = \varnothing$，说明该分布外用户话语只包括未知意图，那么它被称为纯净分布外。多标注分布外检测旨在训练一个打分函数 $S(\mathbf{u})$ 能够同时检测出混合和纯净分布外。

4.2.2　方法总览

如之前所提到的，AIK 旨在识别出测试话语中的所有意图是否都是已知的。形式化上，给定一个测试话语 $\mathbf{u}$，AIK 先预测它的意图数目 r。接着，对于每一个已知意图 $c \in Y_{\text{in}}$，AIK 先提取 $\mathbf{u}$ 中和该意图相关的表示 $\mathbf{v}_c$，接着衡量 $\mathbf{v}_c$ 的概率密度。假设已知意图相关的表示服从条件高斯分布 $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$，这里 $\boldsymbol{\mu}_c$ 是指分布的中心向量，$\boldsymbol{\Sigma}$ 是指协方差矩阵[1]。那么 $\mathbf{v}_c$ 的概率密度可以表示为 $\mathcal{N}(\mathbf{v}_c; \boldsymbol{\mu}_c, \boldsymbol{\Sigma})$，

1　为了计算简便，同已有工作 Yan 等人[99]和 Lee 等人[57]一样，这里我们假设所有已知意图的条件高斯分布有着相同的协方差矩阵。

可以由它与分布 $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$ 的马氏距离 d_c 估算得到[100]：

$$d_c = (\mathbf{v}_c - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}^{-1}(\mathbf{v}_c - \boldsymbol{\mu}_c). \tag{4-1}$$

在为每个已知意图相关的表示计算完马氏距离之后，AIK 对它们取负，并且汇总起来形成 **D**：

$$\mathbf{D} = \{-d_{c_1}, -d_{c_2}, ..., -d_{c_k}\}. \tag{4-2}$$

最终，AIK 使用第 r 大的 **D** 作为 $S(\mathbf{u})$ 来估计话语中是否包含 r 个已知意图。如果一个话语的 $S(\mathbf{u})$ 较低（比如低于预定义的阈值），意味着它包含的已知意图数目小于 r，也就是说，它也包含一个或多个未知意图，因此，可以被分类成分布外。本文把以上过程总结成了Algorithm 3，并在以下段落提供解释。

解释：如果一个话语 **u** 包含已知意图 c，那么 $\mathbf{v}_c$ 将会有着较高的概率密度，即 $\mathcal{N}(\mathbf{v}_c; \boldsymbol{\mu}_c, \boldsymbol{\Sigma})$ 和 $-d_c$ 会比较大，否则 $-d_c$ 会比较小[100]。所以如果 **u** 是分布内，也就是说，**u** 中的所有 r 个意图都是已知的，那么 **D** 中将会有 r 个较大的 $-d$，因此第 r 大的 **D** 也会比较大。如果 **u** 是纯净的或混合的分布外，也就是说，**u** 中的所有 r 个意图不都是已知的，那么 **D** 中较大的 $-d$ 将会小于 r 个，因此第 r 大的 **D** 也会比较小。

尽管 AIK 是从多标注的视角提出来的，它和单标注分布外检测有着很强的联系。在单标注场景中，已有方法往往直接使用最大值来检测分布外，比如最大的 Softmax 概率[56]，最大的 Logit 值[101]，这其实是等价于 AIK 的一种特殊情况，也就是 r 取 1 时。

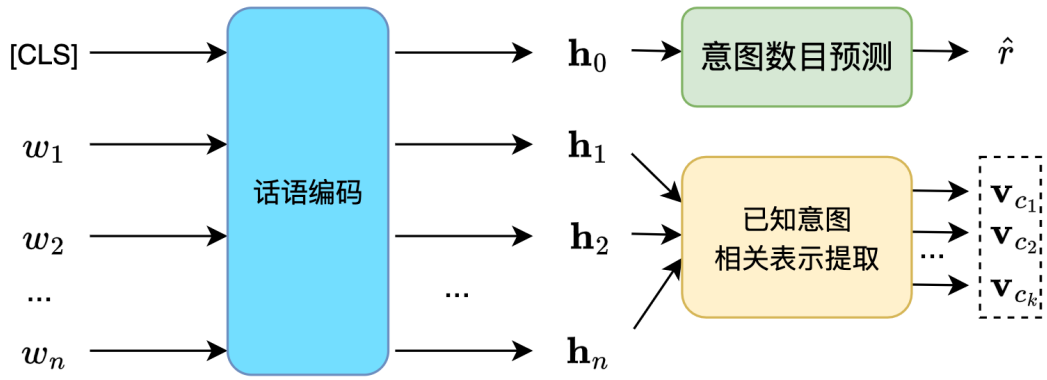

图 4-2 AIK 的模型结构。

4.2.3　模型结构

图 4-2 展示了 AIK 的模型结构，包括话语编码，意图数目预测，已知意图相关表示提取三部分。

话语编码（utterance encoding）。AIK 首先使用了预训练的 BERT[88]来编码话语 $\mathbf{u} = \{w_1, w_2, ..., w_n\}$，这里 n 是话语中的词数目。每个词被编码成了一个固定长度的隐层向量 $\mathbf{h}$，[CLS] 位置的隐层向量被记作 $\mathbf{h}_0$。AIK 使用 BERT 是因为其强大的编码能力，话语的编码器也可以使用其他流行的网络结构，比如 GRU[102]，或 CNN[3]等。

意图数目预测（intent number prediction）。和其他句子级别的任务相似，AIK 采用 $\mathbf{h}_0$ 作为话语整体的表示，并且用它来预测话语 $\mathbf{u}$ 中的意图数目：

$$\hat{r} = f_{mlp}(\mathbf{h}_0), \tag{4-3}$$

这里 $\hat{r}$ 是预测的意图数目, f_{mlp} 是多层感知机网络，能够把 $\mathbf{h}_0$ 映射到一个标量。

已知意图相关表示提取（known intent-wise representation extraction）。受到 Mullenbach 等人的启发，AIK 采用标签级别的注意力机制来得到已知意图相关的表示。具体来说，AIK 先为每个已知意图 c 随机初始化一个可训练的查询（query）向量 $\mathbf{q}_c$，接着使用 $\mathbf{q}_c$ 来计算隐层向量的注意力权重，最终根据计算得到的权重对隐层向量进行加权，得到意图 c 的相关表示 $\mathbf{v}_c$：

$$a_t = \frac{\exp(\mathbf{q}_c^T \mathbf{h}_t)}{\sum_{j=1}^{n} \exp(\mathbf{q}_c^T \mathbf{h}_j)}, \tag{4-4}$$

$$\mathbf{v}_c = \sum_{t=1}^{n} a_t \mathbf{h}_t, \tag{4-5}$$

这里 exp 是以自然常数 e 为底的指数函数。

4.2.4 训练目标

意图数目损失（intent number loss）是模型预测的意图数目和真实意图数目的均方误差 mean-square error, MSE）：

$$\mathcal{L}_{\text{int}} = \mathbb{E}_{(\mathbf{u},\mathbf{y})\sim D}(\hat{r}_{\mathbf{u}} - r_{\mathbf{u}})^2, \tag{4-6}$$

这里 $\hat{r}_{\mathbf{u}}$ 是为 **u** 预测的意图数目，$r_{\mathbf{u}}$ 是真实的意图数目，也就是集合 **y** 的大小。

分布损失（distribution loss）是使得已知意图相关的表示服从可训练的条件高斯分布，对于训练话语中包含的已知意图，AIK 来最大化其概率密度，也就是最小化以下损失：

$$\mathcal{L}_{\text{pos}} = -\mathbb{E}_{(\mathbf{u},\mathbf{y})\sim D}\mathbb{E}_{c\sim\mathbf{y}}\,\mathcal{N}(\mathbf{v}_c; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}). \tag{4-7}$$

对于训练话语中不包含的已知意图，AIK 通过阈值 t 限制其概率密度的大小：

$$\mathcal{L}_{\text{neg}} = -\mathbb{E}_{(\mathbf{u},\mathbf{y})\sim D}\mathbb{E}_{c\sim Y_{\text{in}}\setminus\mathbf{y}} \max(0, t - \mathcal{N}(\mathbf{v}_c; \boldsymbol{\mu}_c, \boldsymbol{\Sigma})). \tag{4-8}$$

总体损失（overall loss）：结合以上的所有损失函数，AIK 的总体损失如下所示：

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{pos}} + \lambda_2\mathcal{L}_{\text{neg}} + \lambda_3\mathcal{L}_{\text{int}}, \tag{4-9}$$

这里 λ_1，λ_2 和 λ_3 是损失权重。

4.3 实验设置

4.3.1 数据集

为了评估 AIK 在多标注分布外检测上的表现，本文基于两个标准的多标注意图分类数据集 MixSNIPS[97]和 MultiWOZ 2.3[104]构造了两个适合分布外检测的数据集。构造细节如下所示：

- MixSNIPS 是 SNIPS 个人语音助手[105]中构造的。我们随机选择两个意图作为验证集的未知意图，另外再选择两个意图作为测试集的未知意图。我

表 4-1　多标注分布外检测数据集 MixSNIPS 和 MultiWOZ 2.3 的统计信息。

Statistic	MixSNIPS	MultiWOZ 2.3
Train-ID	6998	20319
Validation-ID	389	2531
Validation-OOD	664	2236
Test-ID	398	2530
Test-OOD	671	2418
Test-Mixed OOD	489	64
Test-Pure OOD	182	2354
Number of known intents	3	52
Average intent number per utterance	1.6	1.5

们共进行了五种不同的分割。

- MultiWOZ 2.3[104]涵盖了八个不同领域的超过 10K 个对话。对于该数据集，我们随机选择两个领域的意图作为验证集的未知意图，另外再选择两个领域的意图作为测试集的未知意图。我们同样进行了五种不同的分割。

表 4-1 提供了两个数据集五个分割实验的平均总结统计数据。Test-Mixed OOD 是测试集中混合分布外句子，Test-Pure OOD 是测试集中纯净分布外句子。

4.3.2　基线方法

由于单标注和多标注场景有着相同的目标：检测分布外样本，单标注场景下一些有竞争力的方法也可以作为多标注场景下的基线。本文将 AIK 与生成式方法 **Likelihood**，**Likelihood Ratio**（LLR）[51,106]，还有基于分类器的方法 **Energy**[48,58], **Logit**[101], and **LOF**[50,107]进行对比：

- Likelihood 基于分布内话语训练一个语言模型，由该模型推出的似然较低的话语被判定为分布外样本。
- LLR 相较于 Likelihood 额外训练了一个语言模型，该模型是基于扰动后的话语训练得到，最终由两个模型推出的似然比较低的话语被判定为分布外样本。
- Energy 使用二元分类器输出的 Logit 的对数和来检测分布外样本，能量较大的话语被判定为分布外样本。
- Logit 使用二元分类器输出的最大值 Logit 来检测分布外样本，Logit 较低的话语被判定为分布外样本。

- LOF（local outlier factor）基于二元分类器推出的话语表示来检测分布外样本。

4.3.3 实现细节

为了保证公平对比，AIK 和所有基于分类器的方法都使用了预训练的 BERT[88]，本文同时为 Energy 和 Logit 的模型装备了标签级别的注意力机制。本文根据验证集的 AUROC 指标来挑选超参。对于 LOF，本文把最近的邻居数设置为 20。对于 LLR，本文按照 Gangal 等人[106]的建议，使用 UNIGRAM 来引入噪音，并且把噪音的比例设置为 0.5。

对于 AIK，本文只是简单地把 $\lambda_1, \lambda_2, \lambda_3$ 设置为了 1，τ 根据 FPR95 来调整。训练时，本文随机初始化已知意图的表示，为了减小优化难度，并且把协方差矩阵设置为了单位矩阵，此时，就可以直接优化已知意图的表示和已知意图的中心向量的欧氏距离来优化概率密度，阈值 t 被设置成了 300。测试时，本文按照 Lee 等人[57]的建议来计算经验性的均值和协方差矩阵。本文同时使用了四舍五入来保证预测的意图数目是整数。

对于所有的方法和所有的数据集，本文使用了 5 个种子 {0, 1, 2, 3, 4} 来进行测试。由于每个数据集有 5 种不同的分割，结果由 25 次实验平均得到。

4.4 结果和分析

4.4.1 主结果分析

表 4-2 和表 4-3 展示了不同方法在多标注分布外检测上的表现，所有的结果都是百分比。加粗的结果是最好的结果。从结果可知：

- AIK 可以在所有数据集和所有指标上达到最好的效果，并且比基线方法显著要好（根据 t-检验结果，$p\text{-}value < 0.01$）。具体说明，比起最好的基线方法，在 MixSNIPS 数据集上，AIK 可以显著降低 FPR95 指标 15.29%，在 MultiWOZ 2.3 数据集上，可以降低 12.25%。图 4-3 进一步提供了不同方法的 ROC 曲线进行直观的说明。为了可视化方便，本文为每个数据集选择了一种分割。ROC 曲线展示了 AIK 在不同的阈值下，总要比其他方法表现要好。

表 4-2　不同方法在 MixSNIPS 数据集上 AUROC, FPR95, AUPR In, AUPR Out 的表现。

Methods	MixSNIPS			
	AUROC ↑	**FPR95↓**	**AUPR In↑**	**AUPR Out↑**
Likelihood	87.29	55.12	91.17	83.73
LLR	89.40	45.54	92.96	82.73
Energy	68.85	84.66	55.27	78.56
Logit	67.83	84.16	54.81	77.41
LOF	92.79	30.73	88.84	95.14
AIK	**96.29**	**15.44**	**94.93**	**97.46**

表 4-3　不同方法在 MultiWOZ 2.3 数据集上 AUROC, FPR95, AUPR In, AUPR Out 的表现。

Methods	MultiWOZ 2.3			
	AUROC ↑	**FPR95↓**	**AUPR In↑**	**AUPR Out↑**
Likelihood	89.52	76.44	89.18	90.73
LLR	85.90	54.31	85.01	85.71
Energy	89.25	44.31	88.80	89.35
Logit	89.44	43.99	89.06	89.47
LOF	80.68	72.11	78.50	74.56
AIK	**92.22**	**31.74**	**93.33**	**91.01**

- Logit 在 MixSNIPS 数据集上表现较差。这是由于 MixSNIPS 数据集上大部分分布外数据集都是混合分布外，也就是说，它们同样包含已知意图，这会使得 Logit 操作中的 max 操作不再有效。本文将在第 4.4.2 小节对这一点进行深入讨论。
- LOF 和 AIK 在 MixSNIPS 上比 MuliWOZ 2.3 上的表现要好。请注意这两个方法都是基于话语表示，一个好的表示空间，比如相同类别的表示更紧凑是至关重要的。考虑到 MultiWOZ 2.3 的已知类别数目比较多（如表 4-1 所示），模型更难学到一个好的表示空间。
- Likelihood 和 LLR 在两个数据集上表现得平稳，都能获得不错的结果。由于多标注的用户话语的分布更为复杂，生成式模型在多标注场景下可能会更难以拟合训练分布。

4.4.2　混合和纯净分布外检测上的表现

为了对实验结果有进一步了解，本文分别测试了不同方法在混合和纯净分布外样本上的表现。

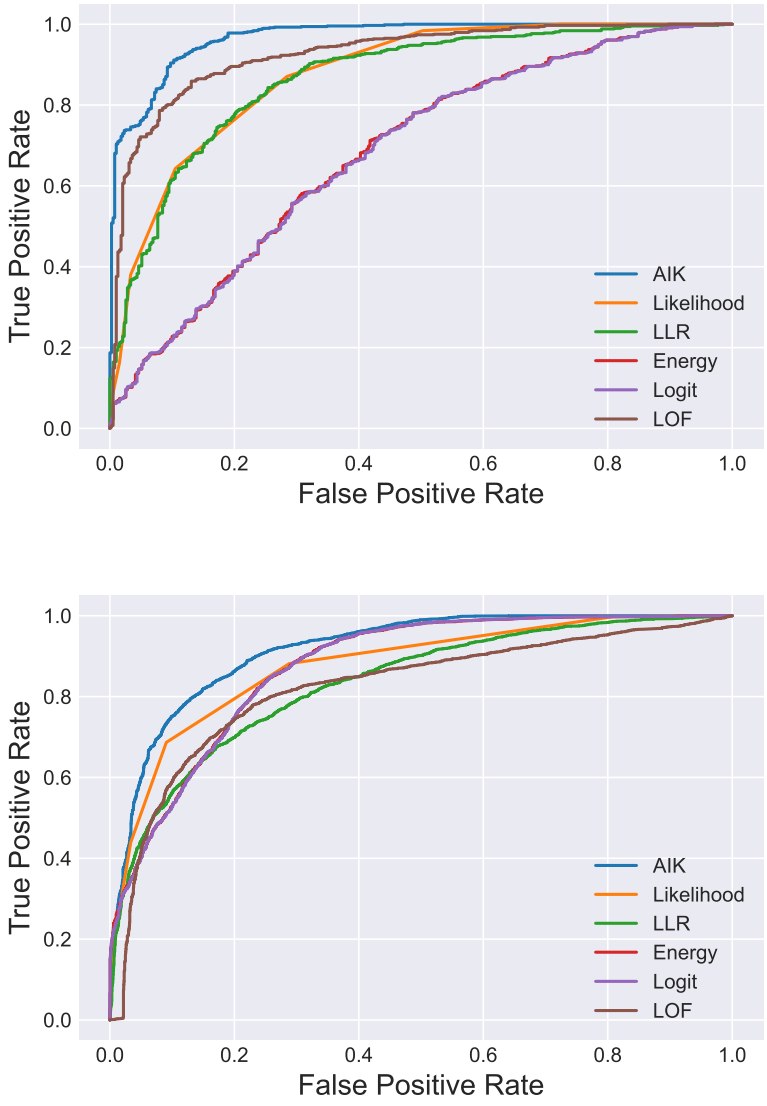

图 4-3 不同方法在 MixSNIPS(上) 和 MultiWOZ 2.3(下) 的 ROC 曲线。

如表 4-4 所示，比起纯净分布外样本，所有方法在混合分布外样本上的表现都要更糟糕，这说明了混合分布外样本更难检测。具体来说，对于 Likelihood 和 LLR，混合分布外样本中的已知意图部分会拉高整个句子的似然；对于 Logit 和 Energy，已知意图部分会导致二元分类器的输出较高。对于 LOF，已知意图部分将会拉近句子的表示和分布内句子表示的距离。这些都会使得混合分布外样本被误分类成分布内样本。

对于 AIK，检测混合分布外样本的表现在 MixSNIPS 数据集上和纯净分布外样本性能接近，但在 MultiWOZ 2.3 上表现出了较大的性能差异。本文将在以下章节揭示这主要是由意图数目预测准确率的不同导致的。

表 4-4　两类分布外样本上的表现。

Method	**MixSNIPS** Pure/Mixed	**MultiWOZ 2.3** Pure/Mixed
Likelihood	95.52/84.19	89.69/83.98
LLR	93.21/87.96	86.44/67.72
Energy	93.37/59.76	89.55/79.10
Logit	92.47/58.69	89.73/79.25
LOF	93.09/92.66	80.78/75.84
AIK	**96.84**/**96.06**	**92.34**/**86.15**

表 4-5　意图数目预测的准确率。

Split	**MixSNIPS** IND/OOD	**MultiWOZ 2.3** IND/OOD
Train	99.46/-	92.79/-
Validation	99.41/85.82	91.83/53.28
Test	99.47/82.70	91.79/53.87

4.4.3　意图数目预测分析

AIK 的表现依赖于意图数目预测的准确率。如第 4.2.2 小节所示，如果一个分布内用户输入有 r 个意图，那么第 r 大的 **D** 将会比较大。但一旦预测的意图数目 $\hat{r}$ 大于 r，那么 $\hat{r}$ 大的 **D** 就会比较小，使得该输入被误分类成分布外。类似地，如果一个分布外用户输入有 r 个意图，那么第 r 大的 **D** 将会比较小。但一旦预测的意图数目 $\hat{r}$ 小于 r，那么 $\hat{r}$ 大的 **D** 可能就会比较大，使得该输入被误分类成分布内。

表 4-5 展示了 AIK 的意图数量预测的准确率。对于分布内样本，MixSNIPS 和 MultiWOZ 2.3 在验证和测试集都可以获得比较高的准确率，比如都大于 90%。对于分布外样本，MixSNIPS 的准确率依旧较高，但 MultiWOZ 2.3 上的准确率却只有 50% 左右。本文认为这可能是由于 MixSNIPS 是手工构造的，句中有明显的特征来指示意图的数量，比如连词“and”。但对于更复杂的 MultiWOZ 2.3 数据集，预测意图数量就不再那么容易。

本文也使用了 golden 的意图数目来测试 AIK 方法的上界，具体来说，是使用 golden 的意图数目来代替预测的意图数目，如表 4-6 所示，为 AIK 装配上 golden 意图数目将会进一步提高分布外样本的检测能力，尤其是在意图数目预测准确率较低的 MultiWOZ 2.3 数据集上。

表 4-6 使用 golden 意图数目在 AIK 上的效果，值是 AUROC。

Method	**MixSNIPS** Pure/Mixed	**MultiWOZ 2.3** Pure/Mixed
AIK	96.84/96.06	92.34/86.15
AIK with Golden intent number	**96.94**/**97.26**	**93.66**/**94.46**

表 4-7 损失函数的消融实验，值是 AUROC。

	$\mathcal{L}_{\text{pos}}$	$\mathcal{L}_{\text{neg}}$	$\mathcal{L}_{\text{int}}$	**MixSNIPS** Pure/Mixed	**MultiWOZ 2.3** Pure/Mixed
1	√			96.22/79.39	92.04/85.25
2	√	√		**97.02**/82.42	**92.86**/85.88
3	√		√	95.14/94.72	91.19/85.83
4	√	√	√	96.84/**96.06**	92.34/**86.15**

4.4.4 消融实验

本文进行了消融实验来调查 AIK 中不同损失的作用。表 4-7 展示了相关的 AUROC 值。

$\mathcal{L}_{\text{neg}}$ 的影响。由表 4-7 可以观察到，$\mathcal{L}_{\text{neg}}$ 带来了更好的结果（第 2 行与第 1 行的对比，第 3 行与第 4 行的对比）。这是由于 $\mathcal{L}_{\text{pos}}$ 损失不能优化类间表示的离散度（inter-class dispersion），也就是对于话语中不包括的已知意图，不能使得与它相关表示的概率密度变小，导致分布内和分布外样本的意图相关表示得不到区分。$\mathcal{L}_{\text{neg}}$ 通过降低相应的概率密度，来提升类间的离散度，进而弥补了 $\mathcal{L}_{\text{pos}}$ 的不足。

$\mathcal{L}_{\text{int}}$ 的影响。在不装配该损失的情况下，AIK 直接使用最大的 **D**，而不是第 r 大的 **D** 作为打分函数。从表中可以看到，$\mathcal{L}_{\text{int}}$ 对混合分布外样本的检测效果有所增益，但却对于纯净分布外样本的检测增益并不明显，甚至造成了性能的轻微下降（第 3 行与第 1 行的对比，第 2 行与第 4 行的对比）。这是由于对于纯净分布外样本，其最大 **D** 就已经能够小于给定的阈值了，所以最大化操作对检测它们就已经足够。然而，对于混合分布外样本，由于它们中包含已知意图，它们的最大 **D** 依旧会非常大，使用最大化操作会使得它们被误分类成分布内样本。

4.4.5 考虑意图之间关系的中心初始化

对于 AIK，在训练的初始阶段，本文随机初始化每个类别的中心向量。但是，对于一些复杂的数据集，比如 MultiWOZ 2.3，随机初始化中心向量可能会

表 4-8　装配 HM 中心的 AIK 在 MultiWOZ 2.3 上的表现。值是 AUROC。

Method	AUROC↑
AIK	92.22
AIK with HM	**94.31**

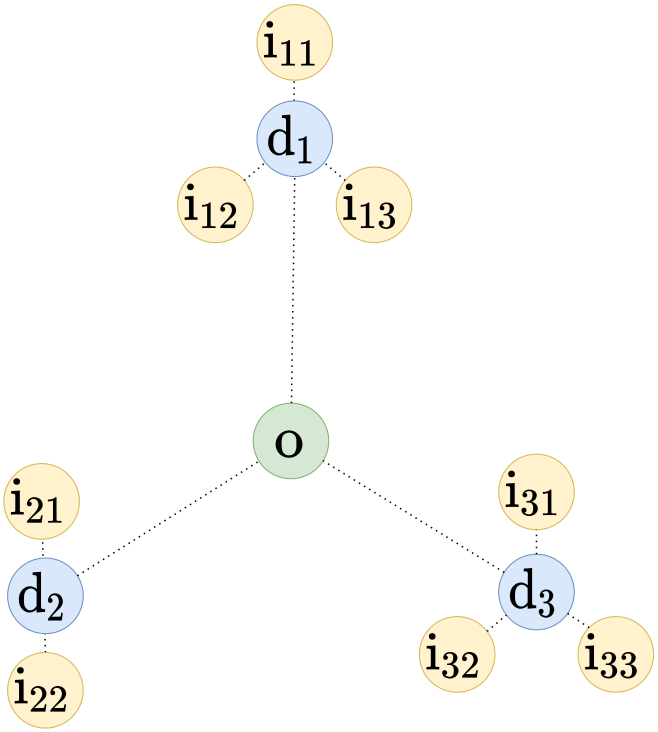

图 4-4　HM 中心的样例。

忽视意图之间的联系，比如对于属于同一个领域的意图，我们期望它们在空间中的位置更近一些。

为了实现这一目标，本文受 Pang 等人[108]的启发使用分层的最大马氏距离中心（Hierarchical Max-Mahalanobis，简称 HM）来初始化每个类别的中心。HM 中心根据类别的树形结构来初始化类别中心。在意图场景下，树形结构是根-领域-意图。具体说明，本文首先为根节点生成中心，使用的是原点。接着，本文在原点周围为每个领域节点生成领域中心。最后，对于每一个领域的附近，本文为该领域下的意图生成意图中心。图 4-4 显示了一个简单例子，以可视化生成的意图中心，这里 o 是根节点，d_m 是第 m 个领域节点，i_{mn} 是 d_m 领域下的第 n 个意图。

如表 4-8 所示，在考虑意图之间的关系后，AIK 的表现可以进一步提升。本文中只考虑了意图之间领域的相关性，考虑意图之间更多的相关性，比如共现性，也将是非常有趣的研究方向。

表 4-9 装配 Logit 的 AIK 在 MixSNIPS 上的表现。值是 AUROC。

Method	AUROC↑
AIK	**96.29**
AIK with logit	95.21

4.4.6 使用 Logit 来核查已知意图数目

在 AIK 中，使用了概率密度，更具体地说是第 r 大的负马氏距离，来核查话语是否包含相同数量的已知意图。作为扩展，我们探索了使用 Logit 核查的有效性。具体说明，本文在基线方法 Logit 中添加了意图数量预测，并使用第 r 大的二元分类器输出来检测分布外样本。理想情况下，较小的第 r 大输出表示其包含的已知意图数量少于其意图数量，可以被分类为分布外样本。

表 4-9 显示了使用 Logit 后，AIK 的性能出现了下降。这是因为 Logit 会受到过自信的影响：对于分布外的样本仍赋予较高的 Logit 值，不如马氏距离可靠[57]。

4.5 总结

本文提出了一个有价值且实用的研究任务——多标注分布外检测。该任务旨在针对用户输入可能包含多个意图的场景，进行分布外用户话语的检测。本文进一步提出了一种新颖的 AIK 方法，通过识别话语中包含的所有意图是否都是已知的，可以同时检测混合分布外和纯净分布外的用户输入。通过在两个数据集上进行实验，本文验证了所提出方法的有效性。同时，本文还通过大量实验对该任务下的新挑战，即检测混合分布外话语，进行了深入分析。

本文的工作仍有以下改进的空间可供探索：

- 对检测到的混合分布外和纯净分布外用户输入做进一步区分。本文的方法虽可以同时检测出这两类分布外样本，但无法实现对两者的进一步区分。区分的意义在于对话系统应当对混合和纯净分布外做分别处理：对于纯净分布外输入，系统可以直接返回安全回复；对于混合分布外用户输入，由于其仍包含已知意图，所以对话系统在返回安全回复的同时，也应当处理其中的已知意图请求。
- 由实验可知，AIK 方法的局限性在于其对意图数目预测的准确性。考虑到

本文只使用 [CLS] 的隐层状态来作为输入进行预测，未来仍有较多的改进空间，比如可以显示考虑一些意图数量相关的特征，比如句子长度，动词数量等，作为输入进行预测。

本工作已总结成文：

- **Yawen Ouyang**, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. "Towards Multi-label Unknown Intent Detection," in *Proceedings of the 29th International Conference on Computational Linguistics.* (**COLING 2022, CCF B 类会议**)

第五章　无标注场景下的分布外检测

5.1　引言

尽管依赖于标注的分布外检测方法取得了不俗的成功，但近年来无标注场景下的分布外检测方法越发受到研究人员的青睐。顾名思义，无标注场景下的方法在训练检测器时不需要依赖于样本标注。这种方法为一些标注昂贵的对话场景（例如医疗、法律、金融等领域）提供了另一种检测分布外样本的可替代方案。

无标注场景下，基于生成式模型的分布外检测方法受到了广泛的关注。图 2-3 展示了基于生成式模型的检测流程。简要总结，在训练时，基于生成式模型的分布外检测方法通过优化生成式模型来拟合训练集用户话语的分布[54]，在测试时，利用训练完成的生成式模型计算用户输入的似然用作检测，似然较低 (小于预设的阈值) 的输入被视为分布外样本，似然较高的则被视为分布内样本。

然而，有研究者发现，完全基于生成式模型所计算出的似然在用于分布外检测存在失效的风险，即分布外样本在测试阶段的似然可能较高，而分布外样本的似然可能较低[54]。在计算机视觉领域，不少工作给出了关于分布外样本较高似然的原因[51,55,109]，认为图片中的背景像素[51]，图片的复杂度[55]，图片的低级特征[109]都会导致生成式模型的失效。在自然语言处理领域，针对文本数据，模型失效的原因探究较少，Gangal 等人[110]虽指出生成式模型失效的原因是受到了文本“表层特征”(surface-level) 的影响，但对“表层特征”并没有给予进一步解释。因此生成式模型在文本的分布外检测为何失效，以及如何缓解，仍是一个亟待解决的问题。

针对基于似然的分布外检测方案失效的现象，本文通过分析生成式模型在检测失效时的样例（似然较低的分布内用户输入和似然较高的分布外用户输入），将失败原因归结为以下两个方面：1). 模型层面，分布内文本建模效果不佳：生成式模型自身对分布内文本的建模仍有提升的空间，即似然可以更高; 2). 数据

层面，输入文本边缘似然的偏差：在对文本边缘似然量化后，本文发现模型输出的似然与文本的边缘似然有着强相关性——对于边缘似然较高的文本，生成式模型同样会输出较高的似然，即使它来自分布外; 对于边缘似然较低的文本，生成式模型同样会输出较低的似然，即使它来自分布内。

为解决上述问题，本文提出了一种新的基于生成式模型的分布外文本检测方法 (记作 Pobe)。具体来说，测试阶段，Pobe 引入了基于 *k*NN 的检索技术来检索训练集中的相似样本，并将由生成式模型对文本的“预测任务”拓展成“预测任务”和“检索任务”的结合，以提升模型对分布内样本的建模效果，缓解模型层面的问题。随后，为了缓解文本边缘似然偏差对检测分布外的影响，本文设计一种校准策略：先使用预训练语言模型 GPT-2[65]来估计文本的边缘似然，接着使用检索增强后的文本似然和文本边缘似然的比值作为打分函数来检测分布外样本，理论上，本文通过贝叶斯定理证明了两者比值的合理性; 实验上，本文在单标注和多标注分布外检测的数据集上都进行了测试，Pobe 超过了其他的无监督的方法，实现了同时期的最佳性能，代码已开源在https://github.com/Gy915/Pobe。

综上所述，本文的贡献如下：

- 从模型层面和数据层面分析，给出基于似然的生成式模型在检测分布外文本时失败的原因：模型自身对分布内文本的建模效果不佳和输入文本边缘似然的偏差。
- 提出一种新的分布外文本检测方法 Pobe 来提升生成式模型的建模效果并减缓文本边缘似然的偏差，实验验证了 Pobe 在检测分布外文本的有效性。

5.2　相关工作

检索增强的生成式模型。随着神经网络技术的发展，自然语言生成任务往往通过生成式模型来学习记忆训练集中的语言知识。然而，近期研究人员注意到，生成式模型往往受限于自身的学习能力，难以学习记忆训练集中的所有语言知识，特别是复杂训练集 (比如句子较长) 中稀有的语言样式[111-114]，这会导致模型对于分布内样本建模不佳。以语言模型任务为例，对于训练集中出现次数较少的专有名词，如“糖醋里脊”，模型在预测阶段，难以通过“糖醋里”这一前缀预测出下个字“脊”，即赋予“脊”字较低的似然。为了解决这一挑战，一个

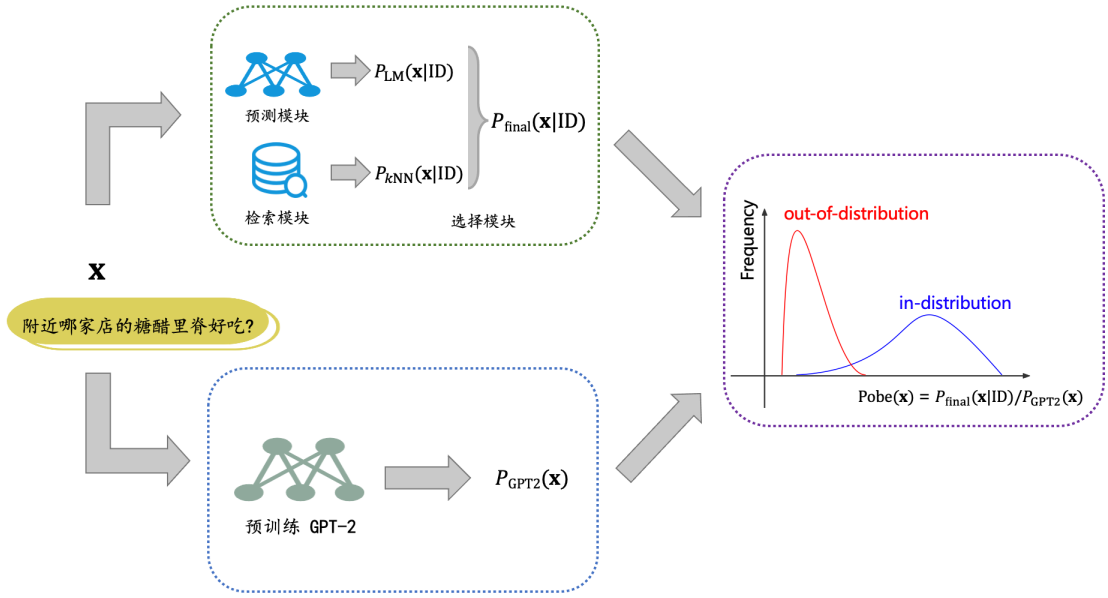

图 5-1 基于 Pobe 的 OOD 检测框架图。

主流的方案是将训练集中的语言知识向量化，并存储至数据库中，在预测阶段从数据库中检索出与预测相关的语言知识用以辅助预测，这种技术被称为检索增强的生成式技术。该技术在语言模型[111]、机器翻译[115-120]、问答[121]领域都有应用。以机器翻译为例，Khandelwal 等人[112]提出的 *k*NN-MT 模型可以将训练集中的翻译知识以向量的形式存储至数据库中，在测试阶段，模型通过向量检索的方式从数据库中获取与当前解码位置最相关的 k 个单词，并使用距离加权融合的方式获取这些单词的分布，最终联合生成式模型的输出作为解码结果。实验显示，检索增强的生成式模型能够有效提升翻译结果。本文受到这些领域工作的启发，首次将检索增强应用在分布外检测任务上，来提升模型对分布内样本的建模能力。

5.3 本文方法

本文提出了 Pobe 来增强生成式模型对分布内用户输入的建模能力和减缓文本边缘似然偏差。具体而言，首先通过引入 *k*NN 检索来计算建模能力增强后的似然 (见第 5.3.1 小节)，其次通过预训练语言模型估计文本边缘似然 (见第 5.3.2 小节)，最终使用两者的比值进行分布外检测 (见第 5.3.3 小节)，整体流程可参考图 5-1 。

需要注意的是，这些过程都只发生于测试阶段，即检测分布外样本阶段，在

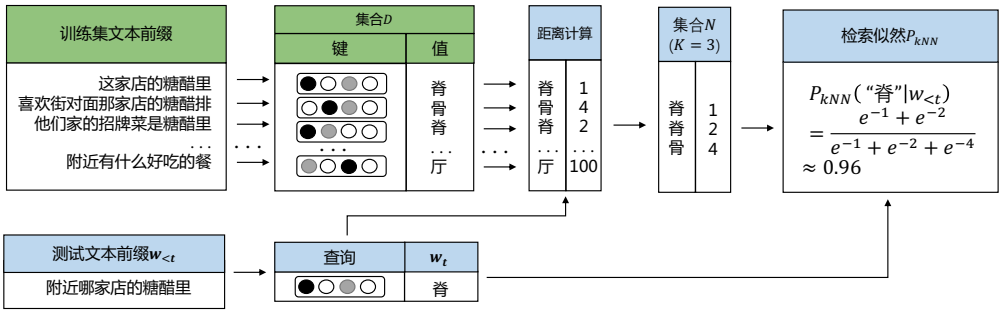

图 5-2　检索模块示例图。

训练阶段不会引入额外的开销，与一般的生成式方法相同，Pobe 的生成式模型的网络结构可基于流行的 GRU 或 Transformer[2]，训练目标可通过对训练集 $\mathcal{X}_{train} = \{\mathbf{x}^i\}_{i=1}^{N}$ 的文本进行极大似然估计，从而得到生成式模型的参数：

$$\theta = \text{argmax}_{\theta^*} \sum_{\mathbf{x}^i \in \mathcal{X}_{train}} \log\, p(\mathbf{x}^i; \theta^*). \tag{5-1}$$

5.3.1　基于 kNN 的检索增强

为了更好地建模分布内样本的分布，在测试阶段，本文在生成式模型预测的基础上，引入基于 kNN 的检索方法，并启发式地选择模型预测结果和检索结果作为文本似然的最终估值。引入检索的动机是，一般的生成式模型通过神经网络参数对训练数据集进行隐式化的记忆，这种方式对分布内样本的建模能力有限，尤其在一些训练集中的稀有的语言样式上[111,122]。而检索式方法对训练数据集进行显示化存储，能够改善隐式化记忆导致的建模能力受限的问题。整个流程可归纳为三个模块，给定需要估计似然的测试文本 $\mathbf{x} = \{w_1, w_2, ..., w_n\}$，其中 w_i 可表示 $\mathbf{x}$ 中的第 i 个字 (也可以表示词)，预测模块通过生成式模型来估计似然，记作 $p_{\text{LM}}(\mathbf{x}|\text{ID})$; 检索模块通过 kNN 来检索训练集中相似前缀来估计似然，记作 $p_{k\text{NN}}(\mathbf{x}|\text{ID})$; 汇总模块以加权的方式汇总两个模块的结果作为文本似然的最终估值 $p_{\text{final}}(\mathbf{x}|\text{ID})$。

预测模块。通过生成式模型来直接估计 **x** 的似然：

$$\log p_{\mathrm{LM}}(\mathbf{x}|\mathrm{ID}) = \sum_{t=1}^{n} p_{\mathrm{LM}}(w_t|w_{<t}, \mathrm{ID}) \tag{5-2}$$

这里 $w_{<t} = \{w_1, w_2, ..., w_{t-1}\}$。

检索模块。通过检索的方式估计 **x** 的似然，该模块需要对训练文本的每个字与其前缀表示进行存储。具体而言，在训练结束后，先对训练集中的每个文本 $x^i = \{w_1^i, w_2^i, ..., w_m^i\}$，使用 θ 对其编码，获取每个字 w_{t-1}^i 的隐层向量作为下个字 w_t^i 的前缀表示[1]，接着将表示作为键 (key)，作为值 (value)，以键值对的形式存储至集合 **D** 中。例如，对于“这家店的糖醋里脊很美味”的训练文本，图 5-2 中绿色模块的第一行以“脊”字为例，在获取其前缀“这家店的糖醋里”的表示，即“里”字的隐层向量后，将该向量和“脊”字以键值对的形式进行存储，用于测试阶段检索。

在检索时，先通过 θ 对测试文本 **x** 进行编码，获得 **x** 中每个字 w_t 的前缀表示，即 w_{t-1} 的隐层向量，接着将该向量用作查询 (query)，计算其与 **D** 中键的欧氏距离，并挑出最近的 K 个键，将这些键对应的值 $\{v_1, ..., v_k\}$，连同各自与查询的距离 $\{d_1, ..., d_k\}$，构成集合 $\mathbf{N} = \{(v_1, d_1), ..., (v_k, d_k)\}$，最后对集合 **N** 中的值进行加权得到 $p_{k\mathrm{NN}}$:

$$p_{k\mathrm{NN}} = \frac{\sum_{(v_i,d_i)\in\mathbf{N}} I_{w_t=v_i} \exp(-d_i)}{Z} \tag{5-3}$$

这里 I 是指示函数，当 $w_t = v_i$ 时返回 1，否则返回 0，$Z = \sum_{(v_i,d_i)\in\mathbf{N}} \exp(-d_i)$ 为归一化系数。例如，针对“附近哪家店的糖醋里脊好吃”的测试文本，图 5-2 的蓝色模块展示了以“脊”字前缀“附加哪家店的糖醋里”为例的查询，先计算其表示与集合 **D** 中键的欧氏距离，接着挑选出最近的 3 个键，将这些键对应的值（脊，脊，骨）与距离（1,2,4）存入集合 N 中，最后通过公式 5-3 计算得到“脊”字的似然。

汇总模块。在得到 $p_{\mathrm{LM}}(\mathbf{x}|\mathrm{ID})$ 与 $p_{k\mathrm{NN}}(\mathbf{x}|\mathrm{ID})$ 后，该模块以加权的方式，汇总

1 由于 w_{t-1}^i 的隐向量编码了 $\{w_1^i, w_2^i, ..., w_{t-1}^i\}$ 的信息，可作为 w_t^i 前缀的表示。

两个分布的结果，作为最终对文本的似然估计：

$$\log p_{\text{final}}(\mathbf{x}|\text{ID}) = \sum_{t=1}^{n} \lambda \log p_{\text{LM}}(\mathbf{x}|\text{ID}) + (1-\lambda) \log p_{k\text{NN}}(\mathbf{x}|\text{ID}), \tag{5-4}$$

这里 λ 是 0 到 1 之间的超参。

5.3.2　基于预训练 GPT-2 的边缘似然

即使 $p_{\text{final}}(\mathbf{x}|\text{ID})$ 对分布内样本有更好的建模效果，本文认为用其作为打分函数来检测分布外样本仍有不足，因为会受到 $\mathbf{x}$ 边缘似然 $p(\mathbf{x})$ 偏差的影响。由于检索增强后的 $p_{\text{final}}(\mathbf{x}|\text{ID})$ 也只是更好地估计 $p(\mathbf{x}|\text{ID})$，而判定 $\mathbf{x}$ 属于分布外类别的标准是当且仅当 $p(\text{ID}|\mathbf{x}) < 0.5$ [1]。根据贝叶斯定理

$$p(\text{ID}|\mathbf{x}) \propto p(\mathbf{x}|\text{ID})/p(\mathbf{x}) \tag{5-5}$$

由于 $p(\mathbf{x})$ 项的存在，$p(\mathbf{x}|\text{ID})$ 并不正比于 $p(\text{ID}|\mathbf{x})$，例如，给定分布内样本 $\mathbf{x}_1$ 和分布外样本 $\mathbf{x}_2$，由前文可知 $p(\text{ID}|\mathbf{x}_2) < p(\text{ID}|\mathbf{x}_1)$，但假如 $p(\mathbf{x}_1) << p(\mathbf{x}_2)$，比如 $\mathbf{x}_1$ 的文本长度远长于 $\mathbf{x}_2$，考虑到 $p(\mathbf{x}|\text{ID}) \propto p(\text{ID}|\mathbf{x})p(\mathbf{x})$，$p(\mathbf{x}_2|\text{ID})$ 反而可能高于 $p(\mathbf{x}_1|\text{ID})$。综上所述，如果使用 $p(\mathbf{x}|\text{ID})$ 检测分布外样本，由于 $p(\mathbf{x})$ 的影响，分布内样本的 $p(\mathbf{x}|\text{ID})$ 可能低于分布外样本，因此直接用 $p(\mathbf{x}|\text{ID})$ 作为打分函数不是最优。

本文为缓解 $p(\mathbf{x})$ 的影响，先对其进行估计，具体选用由预训练语言模型 GPT-2 得到的概率

$$\log p_{\text{GPT2}}(\mathbf{x}) = \sum_{t=1}^{n} \log p_{\text{GPT2}}(w_t|w_{<t}) \tag{5-6}$$

来作为 $p(\mathbf{x})$，由于 GPT-2 是通过数百万的文档进行预训练，包含了广泛的领域，本文假设这些文档中既包括了分布内文本还有分布外文本，因此估计 $p(x|\text{ID}, \text{OOD})$ 可等价于 $p(\mathbf{x})$。

1　当 $\mathbf{x}$ 属于分布外时，$p(\text{ID}|\mathbf{x}) < p(\text{OOD}|\mathbf{x})$，结合 $p(\text{ID}|\mathbf{x}) + p(\text{OOD}|\mathbf{x}) = 1$，推出 $p(\text{ID}|\mathbf{x}) < 0.5$。

Algorithm 4 基于 Pobe 的分布外检测

Input: 测试文本 **x**，生成式模型 θ，预训练 GPT-2，集合 **D**，阈值 τ。

根据公式 5-1 计算 **x** 的检索增强似然 $p_{\text{final}}(\mathbf{x}|\text{ID})$
根据公式 5-6 计算 **x** 的边缘似然 $p_{\text{GPT2}}(\mathbf{x})$
根据公式 5-7 计算 Pobe(**x**)
if Pobe(**x**) $< \tau$ **then**
    return 分布外
else
    return 分布内
end if

5.3.3 基于 Pobe 的 OOD 检测

为缓解边缘似然偏差的影响,本文设计了一种新的校准策略: 使用 $p_{\text{final}}(\mathbf{x}|\text{ID})$ 和 $p_{\text{GPT2}}(\mathbf{x})$ 的比值 Pobe(**x**) 作为打分函数来检测分布外，即：

$$\text{Pobe}(\mathbf{x}) = p_{\text{final}}(\mathbf{x}|\text{ID})/p_{\text{GPT2}}(\mathbf{x}) \tag{5-7}$$

根据 2.2 小节可知，$p_{\text{final}}(\mathbf{x}|\text{ID})$ 用于估计 $p(\mathbf{x}|\text{ID})$，p_{GPT2} 用于估计 $p(\mathbf{x})$，由公式 (5) 可知，两项的比值 Pobe(**x**) $\propto p(\text{ID}|\mathbf{x})$，所以当 Pobe(**x**) 小于给定的阈值 τ, 可判定 **x** 是分布外，否则判定是分布内。阈值可根据用户需求选取，例如实验中用到的 FPR95_IN 评价指标是要求阈值召回不低于 95% 的 ID 样本. 具体检测过程如Algorithm 4 所示。

5.4 实验设置

5.4.1 数据集

为了评估 Pobe 在分布外检测上的表现，本文采取了单标注和多标注场景下的对话理解数据集 CLINC150、MixSNIPS 进行测试。值得注意的是，即使这些数据集都是有标注的，但 Pobe 在训练和推理阶段都不使用这些标签。表 5-1 提供了两个数据集统计数据，关于这两个数据集的详细介绍可参考第 3.4.1 小节, 第 4.3.1 小节。

表 5-1　实验所用数据集的统计信息。

Statistics	CLINC150	MixSNIPS
Train-ID	15000	6998
Validation-ID	3000	389
Validation-OOD	100	664
Test-ID	4500	398
Test-OOD	1000	671

5.4.2　基线方法

本文使用了三个无监督场景下非常有竞争力的方法，包括似然（likelihood），困惑度（perplexity），似然比（likelihood ratio）作为基线方法，更多关于细节可查看第 2.3 节。

- likelihood 取由生成式模型计算的测试样本的似然作为打分函数。
- perplexity 取由生成式模型计算的测试样本的困惑度作为打分函数。
- likelihood ratio 取由生成式模型和扰动后的生成式模型计算的测试样本的似然比作为打分函数。

5.4.3　实现细节

对于 Pobe 中的生成式模型，本文参考 Arora 等人[43]的设置，采用 12 层的预训练 GPT-2[65]作为生成式模型并进行调优，早停 (early stop) 设置为 5，设置初始化学习率为 $5 * 10^{-4}$。对于 Pobe 的检索模块，本文采用 Meta(原 Facebook) 公司开发的 Faiss 工具构造索引。为了兼顾检索的效率和准确率，参考 Jiang 等人[118]，实验采用 IVFFLAT 索引，设置簇类个数为 100，搜寻簇类个数为 10，检索过程中设置近邻个数 k 的范围为 $[0, 8, 16, 32, 64, 128, 256, 512, 1024, 2048]$，$\lambda$ 的范围为 $[0.1, 0.5, 0.9]$ 并通过验证集上分布外检测的性能选择最优超参数: 计算不同 k 取值下验证集的 AUROC 性能，选择最好性能的 k 和 λ 作为超参数。对于边缘似然估计部分，采用 12 层的预训练的未调优的 GPT-2 来估计 $p_{\mathrm{GPT2}}(\mathbf{x})$。对于 LLE 方法和 PPL 方法，为了保证实验的公平性，生成式模型同样采用 12 层的调优后的 GPT-2，调优过程与 Pobe 相同; 对于 LLR 的方法，本文参考 Gangal 等人[106]的设置，采用隐藏层为 1 层，隐藏层大小为 300 的自左向右的 LSTM[123]模型作为语言模型，使用 100 维的 Glove[124]对词表示进行初始化，每次扰动 50% 的文本，使其等概率替换成词表里的随机词。

表 5-2 不同方法在 CLINC150 和 MixSNIPS 数据集上 AUROC，FPR95，AUPR In，AUPR Out 的表现。

Methods	AUROC ↑	FPR95↓	AUPR In↑	AUPR Out↑
CLINC150				
Likelihood	84.6	47.6	96.1	49.2
Perplexity	90.7	32.3	97.8	65.9
LLR	90.2	37.1	97.5	66.4
Pobe (ours)	**93.7 ±0.1**	**26.9 ±1.7**	**98.4 ±0.0**	**78.3 ±0.5**
MixSNIPS				
Likelihood	84.4	56.8	77.9	89.6
Perplexity	85.7	47.6	88.5	81.7
LLR	85.9	54.3	85.0	85.7
Pobe (ours)	**90.9 ±2.5**	**37.2 ±10.1**	**86.1 ±3.1**	**94.1 ±1.9**

5.5 结果和分析

5.5.1 主结果分析

主结果由表 5-2 所示，所有的结果都是百分比，加粗的结果是最好的结果，由此可以观察到，

- 相较于其他无标注方法，Pobe 在两个数据集上的所有指标上都能取得结果最好。其中，关于 FPR95 指标，Pobe 在 CLINC150 数据集上，相较于最好的基线模型能降低 5.4%，在 MixSNIPS 数据集上，相较于最好的基线模型能降低 15.3%。值得注意的是，Pobe 方法在 MixSNIPS 上的结果的标准差较大，这是由于 MixSNIPS 的结果是由该数据集的 5 种不同分割平均得到的。
- 两个基线方法 Perplexity 和 LLR 相较于似然（Likelihood）均有所提升，但不及 Pobe 的性能。Perplexity 可以消除句子长度带来的影响，LLR 可以消除背景词带来的影响，这些影响因素都可包含于句子的边缘似然中：句子越长，边缘似然相对较低；边缘似然的计算包括背景词。但边缘似然也包括一些其他的可能影响句子似然的，与分布内外无关的因素，比如句子的流利度等。

5.5.2　消融实验

为了探究 *k*NN 检索增强和预训练 GPT-2 边缘似然校准两个改进分别对性能的影响，本文对 Pobe 方法进行消融实验，结果如表 5-3 所示，由此可以观察到，两个改进对 Pobe 都有提升。其中，GPT-2 的边缘似然校准方法对检测效果有着较大的提升。以下是对两个改进的进一步分析：

- 为了验证 *k*NN 检索增强对分布内样本建模效果的提升，本文计算了在两个数据集上，分布内和分布外样本在引入 *k*NN 检索增强方法前后的词的负平均对数似然（越小代表建模效果越好），如表 5-4 所示，在使用 *k*NN 检索增强方法后，分布内的样本的分数得到了明显降低，分布外样本的分数得到了明显提高，证明了使用 *k*NN 检索方法能够提升模型对分布内样本的建模效果；
- 为了验证 GPT-2 对边缘似然的校准效果，本文在 CLINC150 和 MixSNIPS 数据集上（随机选取了一种分割）分别随机采样了 200 个分布内和分布外样本，可视化它们的 likelihood 与 p_{GPT2}，Pobe 与 p_{GPT2} 的散点图，如图 5-3 所示（likelihood 与 p_{GPT2} 的在 CLINC150 数据集（左上）和 MixSNIPS 数据集下的关系（左下），Pobe 与 p_{GPT2} 的在 CLINC150 数据集（右上）和 MixSNIPS 数据集下的关系（右下）），在两个数据集上，likelihood 都与 p_{GPT2} 都有着强相关性 (皮尔森系数分别为 0.73 和 0.84)，即对于 p_{GPT2} 较高的样本，likelihood 同样较高，即使它来自 OOD，这一现象的理论原因也在方法章节中通过贝叶斯定理进行了论证。而 Pobe 则可以缓解文本边缘似然的偏差（皮尔森系数分别为-0.22 和 0.37），即 OOD 样本的值总是较低，ID 样本的值总是较高，与它对应的 p_{GPT2} 无关，说明边缘似然偏差问题得到了有效缓解。

表 5-3 Pobe 的消融实验。

Methods	AUROC ↑	FPR95↓	AUPR In↑	AUPR Out↑
CLINC150				
Pobe	**93.7**	**26.9**	**98.4**	**78.3**
Pobe w/o. *k* NN	93.2	29.6	98.3	76.7
Pobe w/o. GPT-2	84.6	46.9	96.2	49.2
Pobe w/o. all	84.6	47.6	96.1	49.2
MixSNIPS				
Pobe	**90.9**	**37.2**	**86.1**	**94.1**
Pobe w/o. *k* NN	90.6	41.4	84.5	93.9
Pobe w/o. GPT-2	84.7	55.8	78.5	89.7
Pobe w/o. all	84.4	56.8	77.9	89.6

表 5-4 词的负平均对数似然

方法	分布内	分布外
CLINC150		
w/o. *k*NN	2.14	5.20
with *k*NN	1.66	5.47
MixSNIPS		
w/o. *k*NN	0.90	8.74
with *k*NN	0.88	9.85

5.5.3 超参分析

为了验证不同 k 取值对 OOD 检测性能的影响，本文在 CLINC150 和 MixSNIPS 数据集上，可视化了不同 k 取值下的 Pobe 的 AUROC 值，由图 5-4 所示，对于不同的 k 值，AUROC 变化幅度不大，且都好于不用 *k*NN 的情况（红色虚线），证明了 k 选择的稳健性。

5.6 总结

无标注分布外检测的目标是在不依赖于训练标注的前提下，检测来自训练分布外的用户输入，为达到这一目标，本文提出 Pobe 方法。Pobe 基于生成式模型，并通过检索的方式来提升模型对分布内样本的建模效果，使用预训练语言模型来缓解样本边缘似然偏差。实验表明，Pobe 方法可以极大提升基线模型的性能。

本工作已总结成文：

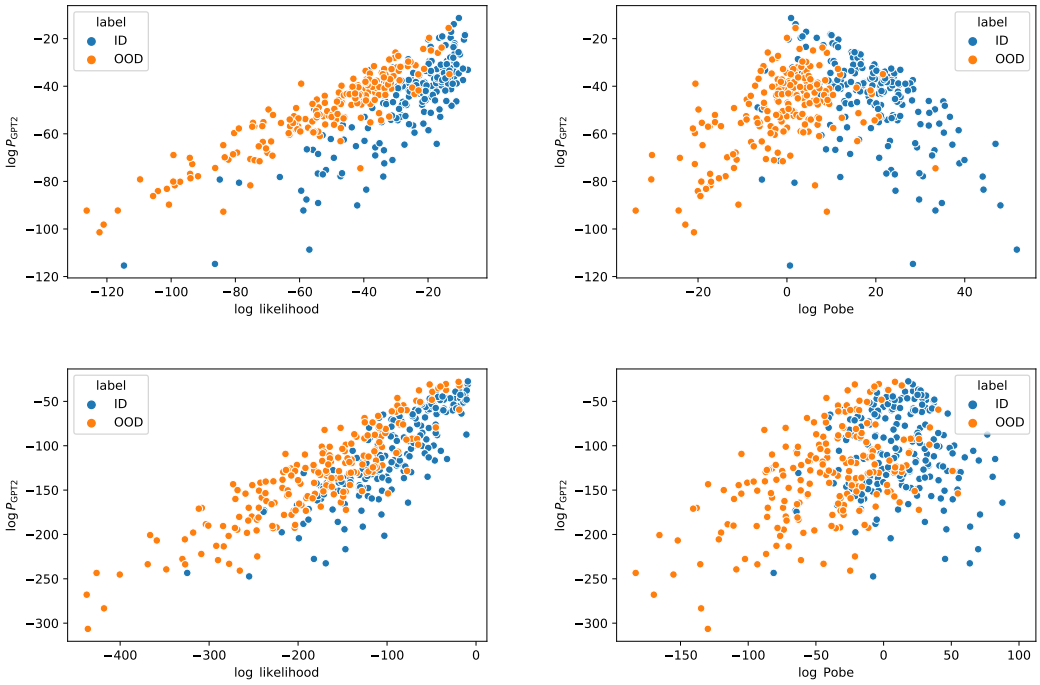

图 5-3　likelihood、Pobe 与 p_{GPT2} 的关系。

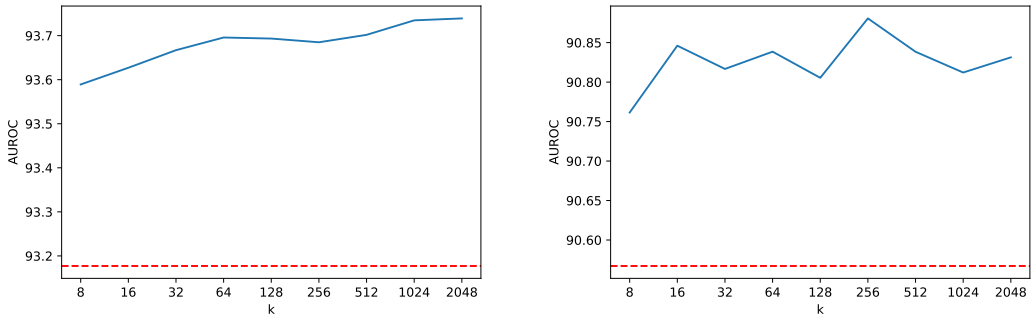

图 5-4　检索模块中近邻个数 k 与 AUROC 之间的关系（CLINC150 在左，MixSNIPS 在右）。

- **Yawen Ouyang**, Yuan Gao, Shi Zong, Yu Bao, and Xinyu Dai. “Pobe, a generative-based out-of-distribution text detection method” in *Journal of Software.* (**软件学报, CCF T1 类期刊**)

与已发表的版本相比，本文做了以下修改：

- 更新了公式 5-4 的计算方式，取得了更好效果。
- 实验上把原数据集替换成了与对话相关的数据集，更符合全文的主题。
- 部分文字和符号的更精准表达。

未来可能的改进方向：

- 引入 *k*NN 检索后，尽管对分布内样本的建模效果有所提升，但会影响模型推理时的速度。为解决这一问题，可在训练生成式模型时，通过蒸馏的方式把 *k*NN 信息加入到训练过程中，进而减弱在推理时对 *k*NN 检索的依赖。
- 设计 ensemble 方法来结合有标注和无标注分布外检测，以实现更强劲的分布外检测器。

第六章　计算资源受限场景下的分布外检测

6.1　引言

不止自然语言处理领域，目前主流的有关分布外检测的研究通常专注于资源丰富的场景，往往忽略了存储检测器带来的资源开销。对于经典的机器学习模型（如支持向量机和逻辑回归）以及经典的神经网络模型（如循环神经网络和卷积神经网络），由于它们的参数量通常较小（小于亿级别），人们往往可以接受训练和存储这些模型所带来的开销。

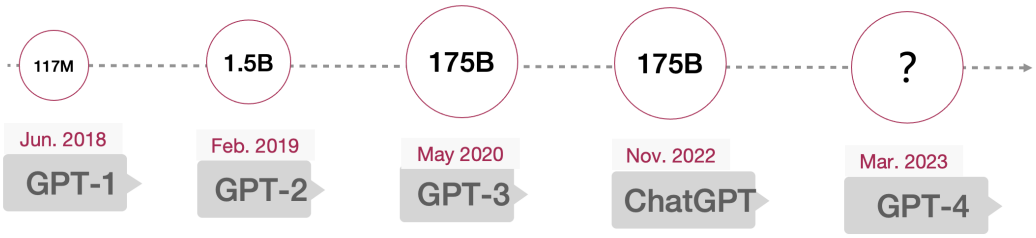

图 6-1　GPT 系列参数量的增长趋势

随着深度神经网络的发展，预训练语言模型已经成为对话领域分布外检测方法的基石[43,125-126]。预训练语言模型的参数量要远超经典模型。以目前比较流行的 GPT-2[65]和 BERT[88]为例，它们的参数量都超过了 1 亿，而 GPT-3[66]，目前最先进的模型之一，它的参数量达到了 1750 亿。关于当前最为先进的 GPT-4 模型，其开发者 OpenAI 公司尚未公开其参数量，但一些研究人员猜测它的参数量已经达到了万亿级别，图 6-1 具体展示了 GPT 系列模型参数量的增长趋势。

随着预训练语言模型参数量的不断增加[127-132]，如果基于流行的调优预训练模型的范式来设计分布外检测器，将会带来巨大的存储开销。在对话场景下，这一问题会愈发明显，因为对话系统的应用场景广泛且更新迭代迅速[133]。如果应用调优范式，我们就需要为不同场景、不同数据集保存一个完整的预训练语言模型，这将极大地浪费资源。这就引出了以下问题：我们可否在不改变预训练语言模型参数的前提下，实现高效的分布外检测？

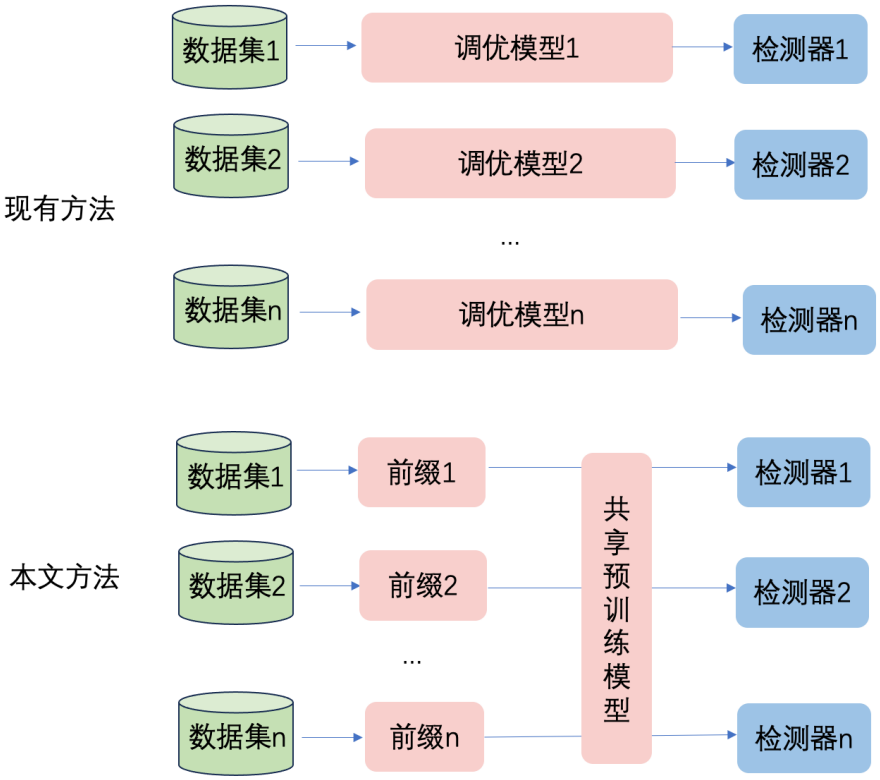

图 6-2　本文方法与之前方法的比较

为了实现这一目标，本文提出了一个无监督的基于前缀调优（unsupervised prefix-tuning）的分布外检测框架（**P**refix-**T**uning based **O**OD detection framework, PTO）。PTO 的核心思想是，给定一个基于训练集语料，使用最大似然训练得到的前缀向量[134]，这里称之为分布内特有前缀（in-distribution specific prefix），预训练语言模型在拼接该向量后，会比不拼接时，赋予分布内文本更高的似然，赋予分布外文本更低的似然。因此本文提出使用由分布内特有前缀带来的似然变化来检测分布外样本：似然提升不明显（如小于预定义的阈值）的为分布外样本。值得注意的是，PTO 的训练过程不需要引入训练语料的标注，这使得它可扩展到标注昂贵的场景。

除了在无标注场景下的应用，本文还对 PTO 进行了扩展，以充分利用训练阶段可用的监督数据。具体而言，本文提出了两个扩展方法，以充分利用训练数据的标签信息和在训练阶段收集到的分布外数据。通过实验证明，这些实用而全面的扩展方法可以进一步提升 PTO 的分布外检测性能。

简而言之，本文提出的 PTO 及其扩展拥有以下优点：（1）**轻量级**：无需调优预训练语言模型的参数，大大节省存储调优后模型所需的开销。图 6-2 直观展示了本文的方法与之前方法的对比，（2）**易于复现**：在训练阶段，只需引入调优

前缀的超参数，而无需引入调优预训练语言模型的超参数。在测试阶段，无需引入其他超参，（3）**理论保障**：在方法章节中进行了详细的论证，为 PTO 及其扩展提供了理论支持。

实验结果显示，PTO 以及其扩展对检测对话场景下的分布外用户输入非常有效。同时，本文也测试了它检测背景分布外样本的性能[43]，以验证它的通用性。对于检测背景分布外样本，PTO 不需要调优预训练语言模型，只需要引入 10M 的前缀向量参数，就超过了基线方法的性能。本文的代码和数据开源在 https://github.com/1250658183/PTO。

本文的贡献总结如下：

- 本文首次探索了计算资源受限场景下的分布外检测，并提出了无需调优预训练模型参数的分布外检测方法 PTO，且不依赖于标注数据。
- 本文为 PTO 方法提供了两个扩展，以充分使用训练数据中可能存在的训练标签和分布外数据，来进一步增强分布外检测性能。
- 实验结果表明，在各种度量指标、分布外检测设置和分布外数据类型下，本文方法的性能与以往基于调优的分布外检测方法表现相当，甚至更好，证明了本文方法的有效性。

6.2 相关工作

分布外检测已经在自然语言处理和计算机视觉领域都得到了持续地关注[43,70,126,135]，一些杰出的无监督方法[43,51,136-137]，标注参与训练的监督方法[58-59,125]，分布外数据参与训练的监督方法[58,61]都被提出。有一些综述对这些方法进行了详尽的总结[53,138]。与之前的工作不同，本文聚焦于资源受限场景下的分布外检测，并提出了不需要调优预训练模型的检测框架，适用于以上多种场景。

前缀调优，作为提示学习（prompt learning）“家族”的一员[139-140]，可以只通过优化一系列拼接在预训练语言模型上的前缀向量，在不调优预训练模型的前提下，触发预训练语言模型生成期望的输出。前缀调优在许多自然语言生成[141-144]和自然语言推理任务[145-147]上都取得了强劲的性能。但它在分布外检测上是否有效还有待探索，据我们所知，本文是第一个全面探索前缀调优在分

布外检测性能的工作。

6.3　本文方法

本章节先介绍了本文提出的方法 PTO（第 6.3.1 小节），接着介绍为 PTO 充分使用训练阶段可利用的监督数据提出的两个扩展（第 6.3.2 小节, 第 6.3.3 小节, 第 6.3.4 小节），最后对它们进行总结（第 6.3.5 小节）。

6.3.1　基于前缀调优的分布外检测（PTO）

PTO 的灵感来自前缀调优：适当的前缀向量可以引导预训练语言模型生成期望的句子[134]，因此我们可以找到特定于训练数据分布的前缀 θ_{in}，以触发语言模型更倾向于生成分布内的句子，即比以前有更高可能性地生成分布内的句子。考虑到预训练语言模型生成所有句子的概率之和（包括分布外和分布外）总是为 1，θ_{in} 会触发语言模型比以前以更低的可能性生成分布外句子。因此，由前缀 θ_{in} 引起的似然提升不明显的，包括下降的，可以视为分布外句子。

具体实现上，本文先按照[134] 的思路，为预训练语言模型的所有隐层前置随机初始化的前缀 θ（本文使用预训练模型是 GPT-2[65]）。接着，本文在保持预训练语言模型参数 θ_{plm} 不变的情况下，通过最大化训练句子的似然来优化前缀：

$$\theta_{in} = \text{argmax}_{\theta} \sum_{\mathbf{x}^i \in \mathcal{X}_{train}} \log\, p(\mathbf{x}^i; \theta, \theta_{plm}). \tag{6-1}$$

这里 $\mathcal{X}_{train}$ 是训练集，$\mathbf{x}^i$ 是训练集中的每个句子。

对 θ_{in} 的训练完成后，本文按如下式子来定义 PTO 的打分函数以检测分布外样本：

$$S_{\text{PTO}}(\mathbf{x}) = \frac{p(\mathbf{x}; \theta_{in}, \theta_{plm})}{p(\mathbf{x}; \theta_{plm})}, \tag{6-2}$$

这里 $p(\mathbf{x}; \theta_{plm})$ 是直接从预训练语言模型中计算得到的 $\mathbf{x}$ 的似然，计算的过程没有前缀向量 θ_{in} 的参与。$S_{\text{PTO}}(\mathbf{x})$ 刻画了前缀带来的似然变化，最终我们可

以借助以下式子来检测分布外样本：

$$G(S(\mathbf{x}),\delta)=\begin{cases}\text{ID} & S_{\text{PTO}}(\mathbf{x})\geqslant\delta,\\ \text{OOD} & S_{\text{PTO}}(\mathbf{x})<\delta,\end{cases} \tag{6-3}$$

关于 $S_{\text{PTO}}(\mathbf{x})$ 的理论洞见：根据贝叶斯规则，$S_{\text{PTO}}(\mathbf{x})$ 与 $p(\text{ID}|\mathbf{x})$ 成正比，如果句子 $\mathbf{x}$ 的 S_{PTO} 比较高，则说明它是 ID 的概率较高。详细解释如下，先使用贝叶斯规则重写 $p(\text{ID}|\mathbf{x})$，我们得到：

$$p(\text{ID}|\mathbf{x})=\frac{p(\mathbf{x}|\text{ID})p(\text{ID})}{p(\mathbf{x})}\propto\frac{p(\mathbf{x}|\text{ID})}{p(\mathbf{x})}. \tag{6-4}$$

本文认为 $p(\mathbf{x};\theta_{plm})$（$S_{\text{PTO}}(\mathbf{x})$ 的分母）可用于估计 $p(\mathbf{x})$，因为预训练模型是在各种大型语料库上进行训练的，不明显偏向于任何特定的领域。为预训练模型添加分布内特有前缀 θ_{in} 后，$p(\mathbf{x};\theta_{in},\theta_{plm})$（$S_{\text{PTO}}(\mathbf{x})$ 的分子）可用于估计 $p(\mathbf{x}|\text{ID})$。因此，它们的比值与 $p(\text{ID}|\mathbf{x})$ 成正比。

6.3.2 装备标签扩展的 PTO（PTO + Label）

在优化 θ_{in} 的过程中，θ_{in} 负责拟合所有训练句子的分布，这无疑会增加 θ_{in} 的训练难度。如果训练数据的标签 $\mathcal{Y}train$ 在训练过程中可以接触到，本章节基于此提出一个新的解决方案：为每个训练标签 y 随机初始化前缀 θ_{in}^{y}，并使用相应标签的句子来优化 θ_{in}^{y}，以便 θ_{in}^{y} 可以专注于拟合标签为 y 的句子分布：

$$\theta_{in}^{y}=\text{argmax}_{\theta}\sum_{\mathbf{x}^{i}\in\mathcal{X}_{train}\wedge\mathbf{y}^{i}=y}\log\ p(\mathbf{x}^{i};\theta,\theta_{plm}). \tag{6-5}$$

对 θ_{in}^{y} 的训练完成后，本文按以下方式来定义 $S_{\text{PTO+Label}}$：

$$S_{\text{PTO+Label}}(\mathbf{x})=\frac{\max_{y}p(\mathbf{x};\theta_{in}^{y},\theta_{plm})}{p(\mathbf{x};\theta_{plm})}. \tag{6-6}$$

$S_{\text{PTO+Label}}(\mathbf{x})$ 的理论洞见：它与 $\max_{y}p(y|\mathbf{x})$ 成比例—高的 $S_{\text{PTO+Label}}(\mathbf{x})$ 表示 $\mathbf{x}$ 有较大的可能性是训练标签之一，即为分布内数据。详细解释如下，为预训练模型拼上标签 y 特定的前缀 θ_{in}^{y} 后，$p(\mathbf{x};\theta_{in}^{y},\theta_{plm})$ 可用于估计 $p(\mathbf{x}|y)$。如前文所

述，$p(\mathbf{x};\theta_{plm})$ 可用于估计 $p(\mathbf{x})$。在假设标签分布是均匀的情况下，$S_{\text{PTO+Label}}(\mathbf{x})$，即 $\max_y p(\mathbf{x}|y)/p(\mathbf{x})$ 的估计值，与 $\max_y p(y|\mathbf{x})$ 成正比。

6.3.3　装备分布外数据扩展的 PTO（PTO + OOD）

如果在训练过程中可以接触到一些分布外数据 $\mathcal{X}_{ood}$，本文对 PTO 提出新的改进以利用它们提高分布外检测的性能。

该扩展的动机是，针对于分布外数据进行调优的前缀 θ_{out}，可以促使预训练语言模型生成分布内句子的可能性变小。因此，由 θ_{in} 和 θ_{out} 同时引起的似然变化，会比只由 θ_{in} 引起的似然变化（也就是 PTO）会更明显。因此，本文使用以下统计量来更新 PTO：

$$S_{\text{PTO+OOD}}(\mathbf{x}) = \frac{p(\mathbf{x};\theta_{in},\theta_{plm})}{p(\mathbf{x};\theta_{out},\theta_{plm})}, \tag{6-7}$$

这里 θ_{out} 是用分布外数据训练得到的前缀：

$$\theta_{out} = \text{argmax}_\theta \sum_{\mathbf{x}^i \in \mathcal{X}_{ood}} \log\, p(\mathbf{x}^i;\theta,\theta_{plm}). \tag{6-8}$$

$S_{\text{PTO+OOD}}(\mathbf{x})$ 的理论洞见：它与 $p(\text{ID}|\mathbf{x})/p(\text{TOOD}|\mathbf{x})$ 成正比，高 $S_{\text{PTO+OOD}}(\mathbf{x})$ 可以被解释为与 TOOD（用于训练的分布外数据）相比，$\mathbf{x}$ 更可能属于分布内。详细解释如下，使用 θ_{out} 作为前缀后，$p(\mathbf{x};\theta_{out},\theta_{plm})$ 可用于估计 $p(\mathbf{x}|\text{TOOD})$。如前文所述，$p(\mathbf{x};\theta_{in},\theta_{plm})$ 可用于估计 $p(\mathbf{x}|\text{ID})$。对 $p(\mathbf{x}|\text{ID})/p(\mathbf{x}|\text{TOOD})$ 变换后，我们得到：

$$\begin{aligned}\frac{p(\mathbf{x}|\text{ID})}{p(\mathbf{x}|\text{TOOD})} &= \frac{p(\mathbf{x}|\text{ID})}{p(\mathbf{x})}\frac{p(\mathbf{x})}{p(\mathbf{x}|\text{TOOD})} \\ &\propto \frac{p(\text{ID}|\mathbf{x})}{p(\text{TOOD}|\mathbf{x})}.\end{aligned} \tag{6-9}$$

6.3.4　同时装备标签和分布外扩展的 PTO（PTO + Label + OOD）

上文提出的两个扩展是不冲突的，如果在训练过程中可以同时使用标签数据和分布外数据，可以同时为 PTO 进行装备：

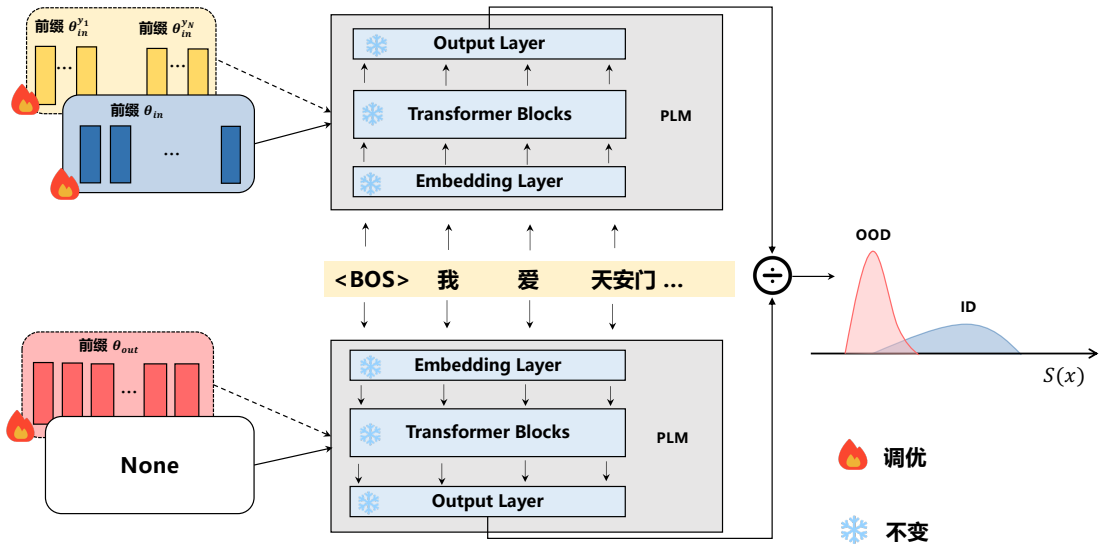

图 6-3 PTO 的示意图。

$$S_{PTO+\text{Label}+\text{OOD}}(\mathbf{x}) = \frac{\max_y p(\mathbf{x}; \theta_{in}^y, \theta_{plm})}{p(\mathbf{x}; \theta_{out}, \theta_{plm})}. \tag{6-10}$$

$S_{\text{PTO}+\text{Label}+\text{OOD}}(\mathbf{x})$ 的理论洞见：结合 $S_{\text{PTO}+\text{Label}}(\mathbf{x})$ 和 $S_{\text{PTO}+\text{OOD}}(\mathbf{x})$ 的证明过程，不难得到 $S_{\text{PTO}+\text{Label}+\text{OOD}}(\mathbf{x})$ 与 $\max_y p(y|\mathbf{x})/p(\text{TOOD}|\mathbf{x})$ 成正比 — 具有较高的 $S_{\text{PTO}+\text{Label}+\text{OOD}}(\mathbf{x})$ 可以被解释为与分布外数据相比，$\mathbf{x}$ 更可能属于训练标签中的一类。

6.3.5 方法总结

PTO 和它的扩展有很多的优势：

- **轻量级**：所有方法都只需要调整和存储少量的前缀向量，而不需要修改预训练模型的参数。
- **易于复现**：除了前缀向量的超参数外（如前缀的长度），所有方法的训练和测试过程不引入任何新的超参数。
- **理论支持**：通过贝叶斯定理的视角，本文提供理论洞察以理解它们的有效性。

一个概述 PTO 及其扩展的示意图如图 6-3 所示，None 是指没有前缀向量拼接到预训练语言模型上。左边的两个虚线框起来的模块表示可替代的扩展：黄色

Algorithm 5 基于 PTO 及其扩展的分布外检测

Input: 训练集 $\mathcal{X}_{train}$，测试样本 **x**.
可选项: 训练集标签 $\mathcal{Y}_{train}$, 分布外数据 $\mathcal{X}_{ood}$.

# 训练过程
if $\mathcal{Y}_{train}$ 不为空 **then**
  for 每个标签 y **do**
    使用公式 6-5 训练 θ_{in}^{y}
  end for
else
  使用公式 6-1 训练 θ_{in}
end if
if $\mathcal{X}_{ood}$ 不为空 **then**
  使用公式 6-8 训练 θ_{out}
end if
# 测试过程
if θ_{out} 和 θ_{in}^{y} 都不可得 **then**
  使用公式 6-2 计算 S_{PTO}
else if 只有 θ_{in}^{y} 可得 **then**
  使用公式 6-6 计算 $S_{\text{PTO+Label}}$
else if 只有 θ_{out} 可得 **then**
  使用公式 6-7 计算 $S_{\text{PTO+OOD}}$
else
  使用公式 6-10 计算 $S_{\text{PTO+Label+OOD}}$
end if

是为标签设计的扩展，红色是为分布外数据设计的扩展。本文也在Algorithm 5 中总结了它们的训练和测试过程。

6.4　实验设置

6.4.1　数据集

本文所使用的数据集包括对话理解数据集 CLINC150 和背景分布外数据集 IMDB-YELP:

- 如前文所述，CLINC150 覆盖了语音助手场景下，用户话语中可能出现的多种意图。分布外数据是具有未知意图的用户话语。它在训练数据集中提供了可用于训练的分布外数据。
- IMDB-Yelp 数据集中 IMDB 作为分布内数据，Yelp 作为分布外数据。IMDB 是一个长电影评论数据集，Yelp 是一个商业评论数据集。因为 IMDB 和 YELP 都没有提供验证集，为了方便实施早停策略，本文从 IMDB 和 Yelp

表 6-1 本章节用到的数据集的统计量。

Statistics	CLINC150	IMDB-Yelp
Train-ID	15000	25000
Train-Label	150	2
Train-OOD	250	-
Validation-ID	3000	10000
Validation-OOD	100	10000
Test-ID	4500	25000
Test-OOD	1000	38000

的未标注数据集中选了 10000 条句子作为验证集。

表 6-1 提供了两个数据集的统计信息。Train-Label 是训练集句子的标签。

6.4.2 基线方法

本文使用了性能强劲的有监督分布外检测方法马氏距离（Mahalanobis distance，简称 Mahalanobis）[57,125]，能量（Energy）和校准能量（Energy + OOD）[48,58], MLS[59] 作为基线方法。基于训练集样本和它的标签训练得到分类器,

- **Mahalanobis** 使用样本表示和类别条件高斯分布的最近马氏距离作为打分函数。
- **Energy** 使用分类器的输出的 Logit 指数和作为打分函数。
- **Energy + OOD** 在训练阶段，使用训练集提供的分布外样本来调整分布内和分布外之间的能量差距。
- **MLS** 使用分类器的输出的最大 Logit 作为打分函数。

本文也使用了有竞争力的无监督方法 IMLM + BCAD + MDF[136], 困惑度（likelihood）[43], 似然比（likelihood ratio，简称 LLR）[51,137]作为基线方法:

- **IMLM + BCAD + MDF** 同样使用马氏距离作为打分函数，同样使用两种领域特定的微调方法来进一步提升性能。
- **likelihood** 使用分布内数据来微调预训练的 GPT-2 模型，并使用困惑度来作为打分函数检测分布外样本。
- **LLR** 使用分布内数据训练了一个正常的从左到右的基于 LSTM 语言模型[123]，并使用扰动的分布内的句子训练一个背景语言模型。似然比使用两个模型的似然的比值来检测分布外样本。

6.4.3　实现细节

对于所有的方法，本文基于验证集上的 AUROC 表现来选择超参，和进行早停策略。

对于 PTO 框架，本文使用了 huggingface 实现的 GPT2-base 作为基础的预训练模型[148]，并且前缀优化的实现基于开源框架 OpenPrompt[149]。所有的结果都是在 5 个不同的种子上平均的。前缀长度对于结果有重要影响，所以本文从 $10, 50, 100, 200, \mathbf{300}, 500$ 中进行搜索。对于 PTO+Label, 总的前缀长度 300，被平均分配到每个标签上。对于 PTO+OOD，分布外的前缀长度也被设为 300。PTO+Label+OOD 的超参数与 PTO+OOD 和 PTO+Label 保持一致。

对于有监督的基线方法，本文使用预训练的 BERT[88]作为编码器，并使用交叉熵损失来优化。对于能量，本文遵循 Liu 等人[58]的设置把温度 $\boldsymbol{T}$ 设置为了 1。句子表示上，本文使用了平均池化来作为聚合方式。

对于无监督的基线方法，IMLM + BCAD + MDF 结果来自作者的开源实现。对于 PPL，本文同样使用 GPT2-base 作为解码器。对于似然比方法，本文遵循 Gangal 等人[137]，并使用一层，300 维的 LSTM，词向量是由 100 维的 Glove 初始化得到[150]。为了训练背景模型，本文对句中的 50% 的词进行扰动，替换成词典中的随机采样的词。

6.5　结果和分析

6.5.1　主结果分析

表 6-2 表 6-3 展示了不同方法在检测分布外样本上的性能，↑ 表示值越高越好，↓ 表示值越低越好。#Params 指的是调优的参数量。每个设置的最好结果都进行了加粗。所有的数字都是百分比。因为 IMDB-Yelp 没有提供可训练的分布外数据，本文只报告了在 CLINC150 数据集上的PTO + Label + OOD 的表现。我们可以观察到：

- **PTO 在所有数据集和指标上都比无监督基线方法表现更好。**对于 CLINC150 数据集，相比最好的无监督基线，PTO 将 FPR95 降低了 **4.5%**，并且在 IMDB-Yelp 上，PTO 也要比基线表现好 **6.4%**。为了进一步了解，我们在图 6-4 中

表 6-2 不同方法在 CLINC150 数据集上的分布外检测性能。

Method	AUROC ↑	FPR95 ↓	AUPR In ↑	AUPR Out ↑	#Params
Unsupervised					
IMLM + BCAD + MDF	83.7 ± 0.4	62.9 ± 1.5	95.3 ± 0.2	54.6 ± 1.8	110M
PPL	90.7 ± 0.3	32.3 ± 2.2	97.8 ± 0.1	65.9 ± 1.2	124M
LLR	90.2 ± 0.3	37.1 ± 1.5	97.5 ± 0.1	66.4 ± 1.3	3.7M
PTO (ours)	**92.8 ± 0.1**	**27.8 ± 0.9**	**98.3 ± 0.1**	**73.8 ± 0.5**	10M
Supervised					
Mahalanobis	97.4 ± 0.1	10.5 ± 0.6	99.4 ± 0.0	89.6 ± 0.6	110M
Energy	97.6 ± 0.0	10.2 ± 0.4	99.4 ± 0.0	92.0 ± 0.3	110M
Energy + OOD	**98.1 ± 0.1**	**8.2 ± 0.6**	**99.5 ± 0.0**	**93.9 ± 0.3**	110M
MLS	97.5 ± 0.1	10.4 ± 0.3	99.4 ± 0.0	91.6 ± 0.3	110M
PTO + Label + OOD (ours)	96.7 ± 0.4	17.6 ± 1.6	99.2 ± 0.1	89.3 ± 0.8	20M

表 6-3 不同方法在 IMDB-YELP 数据集上的分布外检测性能。

Method	AUROC ↑	FPR95 ↓	AUPR In ↑	AUPR Out ↑	#Params
Unsupervised					
IMLM + BCAD + MDF	97.4 ± 0.0	9.2 ± 0.1	97.2 ± 0.0	97.8 ± 0.0	110M
PPL	88.9 ± 0.1	41.7 ± 0.2	85.9 ± 0.2	91.6 ± 0.1	124M
LLR	90.8 ± 0.4	40.5 ± 1.0	87.9 ± 0.4	93.7 ± 0.3	71M
PTO (ours)	**99.3 ± 0.1**	**2.8 ± 0.4**	**99.2 ± 0.1**	**99.6 ± 0.1**	10M
Supervised					
Mahalanobis	97.0 ± 0.2	11.7 ± 2.7	96.4 ± 0.8	97.6 ± 0.5	110M
Energy	76.5 ± 1.2	53.8 ± 2.8	75.6 ± 1.2	77.0 ± 1.6	110M
MLS	76.5 ± 1.3	53.8 ± 2.8	75.5 ± 1.3	77.1 ± 1.2	110M
PTO + Label (ours)	**99.6 ± 0.1**	**2.0 ± 0.2**	**99.4 ± 0.1**	**99.3 ± 0.0**	10M

展示了 PTO 和 PPL 分数的直方图分布。我们可以看到，相比于 PPL，PTO 在 ID 和 OOD 之间的区分度更高，从而实现更有效的 OOD 检测。

- **PTO + Label (+ OOD) 在背景分布外检测上的表现远远优于监督基线，在语义分布外检测上也实现了有竞争力的表现**。值得注意的是，这里所有的监督方法都需要调整预训练语言模型，而本文提出的方法不需要，因此本文的方法在提供了有效性的同时仍然非常轻量（PTO+Label+OOD）只调整了 20M 的参数，不到有监督方法的 20%。本文也把 PTO + Label + OOD 扩展到了 GPT2-medium 的预训练模型下，分布外检测性能可以进一步得到提升（CLINC150 数据集上的 AUROC 可以达到 96.9%）。

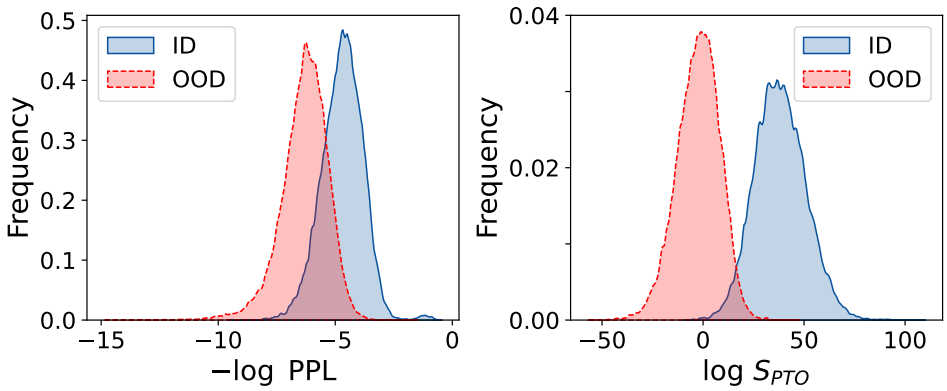

图 6-4　困惑度（左图）和 PTO（右图）在 IMDB-Yelp 数据集上的检测分数的直方图。

表 6-4　PTO和它扩展的性能比较。表中的值是测试集上的 AUROC 指标。

Method	CLINC150	IMDB-Yelp
PTO	92.8 ± 0.1	99.3 ± 0.1
PTO + Label	94.3 ± 0.2	**99.6 ± 0.1**
PTO + OOD	**95.4 ± 0.3**	-

6.5.2　标签扩展的效果

在相同调优参数的数量下，PTO + Label 比 PTO 有着性能增益。正如表 6-4 所示，在 CLINC150 数据集上的增益更加明显，在 AUROC 指标上有着 1.5% 的提升。值得注意的是，它们都有着相同的前缀长度：每个都是 300。

PTO + Label 比起 PTO 可以赋予给分布内句子更高的似然。具体而言，给 PTO 配上标签扩展后，验证集上的分布内句子的平均困惑度，在 CLINC150 数据集的验证集上由 **3.01** 下降到了 **2.23**，在 IMDB-Yelp 数据集上由 **3.72** 下降到了 **3.70**。在 CLINC150 数据集上提升更明显的原因是，它相较于 IMDB-Yelp 有着更多的标签数目（150 比 2）。

比起 PTO，PTO + Label 有着更快的收敛速度。如图 6-5 所示的实验结果表明, PTO + Label 在验证集上的最佳轮数是 **9**, 而PTO的最佳轮数是 **16**。原因是有了标签扩展，每个标签都可以专注于优化自身的前缀。

6.5.3　分布外扩展的效果

PTO + OOD 比 PTO + Label 在 CLINC150 上更有效果。由表 6-4 结果所示，PTO + OOD 在 CLINC150 上的 AUROC 表现比PTO + Label 高 **1.1%**。本文推测，

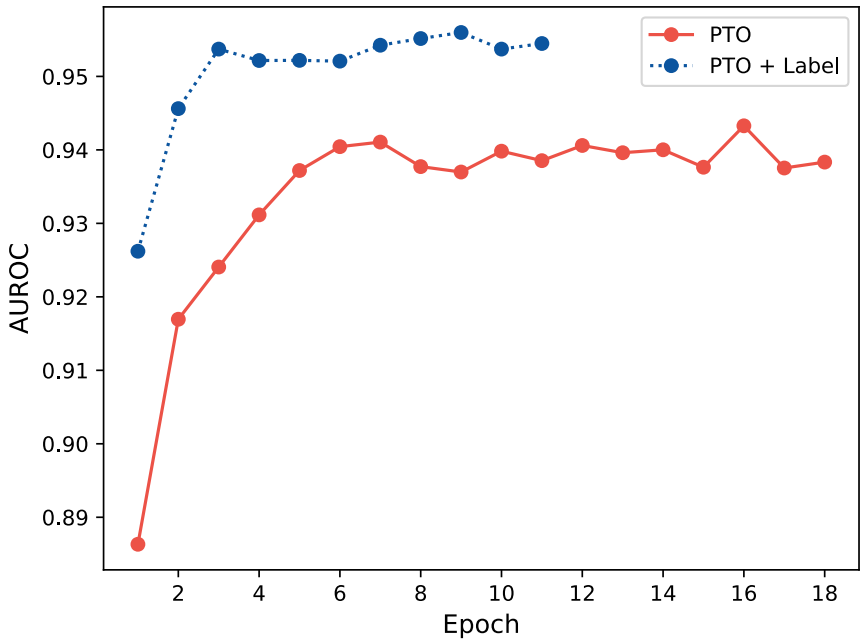

图 6-5 PTO 和 PTO+Label 在不同轮次下，CLINC150 验证集上的 AUROC 表现。

给训练数据配上分布外数据，比起配上标签，可以使得训练数据和测试数据的分布差距更小。

PTO + OOD 仍易于复现。训练 OOD 前缀的超参数与 ID 前缀相同，因此PTO + OOD 不需要任何新的超参数。相比之下，使用校准能量需要进行大量超参数的调整工作，比如 Hinge 损失的两个边界值，损失的权重[58]。

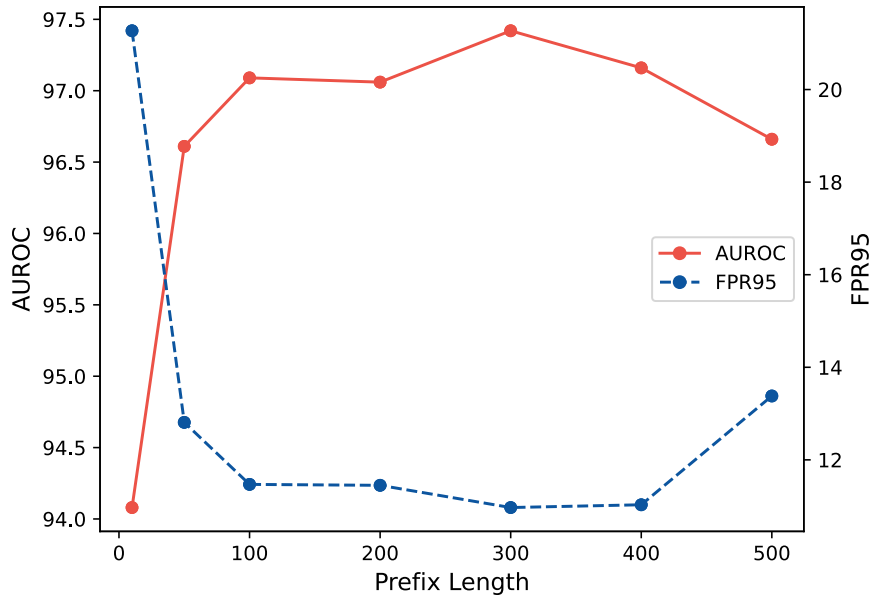

图 6-6 PTO 在不同前缀长度下，IMDB-Yelp 验证集上的 AUROC 和 FPR95 表现。

6.5.4　前缀长度的影响

前缀长度是 PTO 的一个关键超参，之前的工作表明最优的前缀长度因任务而异[134]。受此启发，本文实验了从 10 到 500 的不同前缀长度，来观察它对分布外检测的性能。图 6-6 的结果显示，在这个范围内，性能随着前缀长度的增加而增加，直到 300 后开始下降。本文认为这是合理的，因为较长的前缀长度往往会过拟合训练数据，进一步降低验证集的性能。

表 6-5　分布内和分布外句子开头占比最高的前 10 个二元组（2-gram）和它们的百分比。

分布	2-gram / percent
分布内	can you/6.1, i need/4.8, what is/4.5, what 's/3.6, tell me/3.1, i want/2.0, how do/2.0, how much/1.8, how many/1.8, how long/1.6
分布外	can you/6.6, what is/5.9, what 's/5.3, how many/4, tell me/4, how do/3.6, what are/3.1, how much/2.7, look up/2.1, find out/1.8

6.5.5　错误分析

被 PTO 误分类的分布外句子和分布内句子往往有着相同的前缀词。具体而言，在 CLINC150 上检查未被 PTO 检测到的 OOD 句子时（即具有更高的 S_{PTO}），我们观察到它们在句子开头的前两个词通常可以在分布内句子中找到（见表 6-5，分布内和分布外重叠的二元组被标为了蓝色）。考虑到 $\log S_{PTO}$ 的得分是由句子 $\mathbf{x}$ 中的每个词 w_i 的得分进行加和：

$$\log S_{PTO}(\mathbf{x}) = \sum_{w_i \in \mathbf{x}} \log \frac{p(w_i|\mathbf{w}_{<i}; \theta_{in}, \theta_{plm})}{p(w_i|\mathbf{w}_{<i}; \theta_{plm})}, \tag{6-11}$$

而这两个词会导致分布外句子更高的得分，具体得分如图 6-7 所示。

其根本原因在于 PTO 利用从左到右的 GPT-2 模型来估计句子的似然。在推导句中前些位置的词的似然时，后面位置的词是不可见的。因此，在这种情况下，对于 GPT-2 模型，分布内和分布外之间没有区别，PTO 会像处理分布内样本一样将分布外样本的前些词赋予更高的分数。我们将其解决方案留给未来的工作。

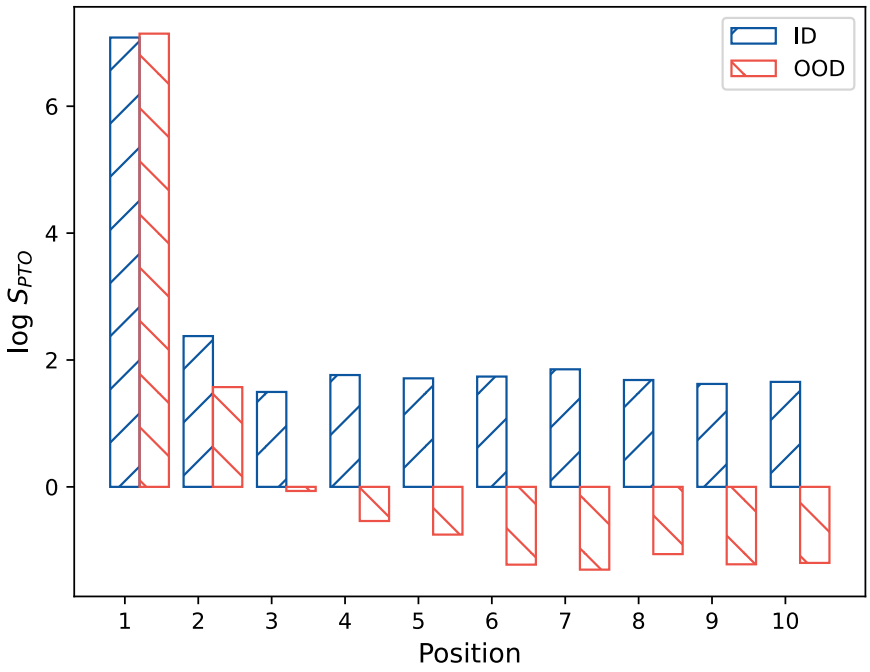

图 6-7 分布内和分布外句子中每个位置词的平均 $\log S_{\text{PTO}}$ 分数。

表 6-6 使用基于前缀优化分类器推导出的能量和 MLS 在检测分布外样本的效果。

Method	AUROC ↑	FPR95↓	AUPR In↑	AUPR Out↑
MLS	92.22	36.95	97.41	78.07
Energy	92.41	33.75	97.57	78.14

6.5.6 基于前缀优化分类器的分布外检测

为了深入研究前缀调整在 OOD 检测中的潜力，我们还在 CLINC150 数据集上进行了基于前缀优化分类器的实验[145,149]。特别地，我们使用意图作为标签词构建手动 verbalizer[151]。同时，本文将原始输入 $\mathbf{x}$ 修改为模板形式 $\mathcal{T}(\mathbf{x}) =$ `[PREFIX]`$\mathbf{x}$`[MASK]`，然后基于 `[MASK]` 成为每个标签词的概率对 $\mathbf{x}$ 进行分类。表 6-6 展示了基于分类器的 Energy 和 MLS 得分的性能。我们可以看到它们的性能不如PTO + Label。本文认为这种策略的局限性在于它依赖于模板和语言表述的设计，而本文提出的方法PTO + Label 不需要它们。

6.5.7 为 PTO 装配更大的预训练模型

为了展示本文提出方法的对于不同的预训练模型的通用性，本文为 PTO+Label+OOD 装配上了 GPT2-medium，并测试了其在 CLINC150 上的表现，由表 6-7 可知，装配上更大的预训练模型后，本文提出方法的性能可以进一步提升，展现了良好

表 6-7　不同预训练模型下的方法性能。

Model	**AUROC ↑**	**FPR95↓**	**AUPR In↑**	**AUPR Out↑**
GPT2-base	96.7 ± 0.4	17.6 ± 1.6	99.2 ± 0.1	89.3 ± 0.8
GPT2-medium	96.9 ± 0.1	14.8 ± 0.8	99.2 ± 0.0	90.0 ± 0.2

的通用性。

6.6　总结

本章节中，我们关注了计算资源受限场景下的分布外检测问题，这在之前的研究工作中被忽略。本文提出了一个无监督的，基于前缀调优的框架 PTO，填补了这一领域的空白。此外，本文扩展了 PTO，充分利用了训练阶段可选的训练标签和分布外句子。本文提出的方法具有轻量级、易于复现和理论上的正当性等关键优势。本文揭示了 PTO 及其扩展在语义和背景分布外检测上的有效性。我们希望我们的工作能够为未来的研究提供有价值的起点，并激发他们探索轻量级分布外检测的更多可能性。

然而，本文工作仍存在以下两个限制，可以作为未来的研究方向：

- 本文提出的轻量级分布外检测框架是基于前缀调优技术设计的。然而，亦可能存在其他技术来实现这个目标，比如 Adapter，LoRA 等，探索并比较不同技术在分布外检测的异同会非常有趣。
- 对于 PTO + Label，每个标签都只优化自身前缀，忽略其他标注的前缀，考虑到不同标签的句子间存在共享的特征，这种零耦合的前缀可能会导致前缀冗余问题，未来可以设计共享前缀来优化标签间共享的句子特征。

本工作已总结成文：

- **Yawen Ouyang**, Yongchang Cao, Yuan Gao, Zhen Wu, Jianbing Zhang and Xinyu Dai. “On Prefix-tuning for Lightweight Out-of-distribution Detection,” in *Proceedings of the 61th annual meeting of the association for computational linguistics.* (**ACL 2023, CCF A 类会议**)

第七章　总结与展望

本文工作主要涉及国家自然科学基金“融合细粒度情感分析的推荐系统研究”和“基于语言认知机理的汉语框架语义计算研究”等科研项目内容。

7.1　本文贡献

尽管有关任务型对话系统的研究在近年来受到了广泛关注，但现有研究工作主要集中在封闭环境下，忽略了开放环境下对分布外用户输入的检测，这对系统的稳健性造成了损害。为此，本文旨在探索面向任务型对话系统的分布外检测，并对单标注、多标注、无标注和计算资源受限场景进行了系统研究。本文的主要创新点如下：

- **单标注场景下的分布外检测**。针对单标注场景，分布外检测器会受到偏差词的影响，本文首先设计了基于梯度的偏差词检测方法，接着基于检索得到的包含相同偏差词的分布外样本，设计了新的训练目标来校准检测器。本文通过实验证明，校准后的检测器在不同打分函数、数据集上都得到了性能提升，并且不会影响对分布内样本的分类性能。
- **多标注场景下的分布外检测**。针对单标注场景假设过强的问题，本文将其扩展到多标注场景，并提出了对混合分布外样本进行检测的新挑战。为应对该挑战，本文提出了一种新的分布外检测方法，可以同时检测混合和纯净分布外样本。实验结果表明，本文提出的方法在多标注场景下取得了最先进的性能，并通过大量的分析证明了检测混合分布外样本的挑战性。
- **无标注场景下的分布外检测**。针对基于生成式模型的无标注检测，往往会受到建模效果不佳和文本边缘似然偏差的问题。本文引入 k 近邻检索来提升建模效果，并引入预训练模型建模文本的边缘似然。实验表明，本文的方法可以大幅度提升生成式模型的检测效果，分析实验也论证了上述两个问题，通过本文的方法得到了缓解。

- **计算资源受限场景下的分布外检测**。针对现有检测器都需要调优和存储预训练模型的问题。本文提出了一个无监督的，基于前缀调优的分布外检测框架来解决该问题，本文提出的框架不需要调优大模型，大大节省资源开销。同时，本文也为该框架提供了扩展，使得其更适用于有监督的分布外检测场景。实验表明，本文在不调优大模型的前提下，可获得相比调优的方法有竞争力的结果。

综上所述，本文的研究对任务型对话系统下的分布外检测进行了全面而深入的探索，填补了相关研究领域的空白。与已有的工作相比，本文具有以下显著特点和优势：

- **系统性**：本文在不同场景下系统地研究了任务型对话系统中的分布外检测问题，深入分析了其中的挑战，并提出了相应的解决方案。
- **实用性**：本文首次将对话场景中的分布外检测任务扩展到多标注场景，使得模型能够同时检测多标注情况下的分布外数据，更贴近真实场景。此外，本文还研究了计算资源受限的场景下的分布外检测，提出了针对有限存储资源情况下的解决方案。这些实用性的研究扩展了分布外检测的应用范围。
- **先进性**：本文的研究在不同场景下取得了同期最好或有竞争力的结果，并都发表于知名会议，体现了本文提出方法的先进性。

7.2 未来方向

虽然本文在不同场景下对任务型对话系统下的分布外检测问题进行了系统研究，并取得了明显的改进，但本文认为该领域仍存在以下需要进一步研究的问题：

- **多模态场景下的分布外检测**。多模态相关的研究已经得到广泛关注[152-156]。在对话场景中，用户的输入可以包含文本、语音、图像等多种模态信息，因此多模态对话系统具有广泛的应用场景。在这种设置下，如何对齐不同模态的特征空间、消除某一模态与意图标签之间的伪相关性，是一个需要进一步思考和研究的问题。
- **大语言模型下的分布外检测**。尽管大型模型相较于一般模型具备更强的能力，但它们同样受制于认知边界，尤其是在特定领域的大型模型中。因此，

对于大型模型而言，分布外检测显得尤为重要。在技术发展方向上，无论是为大型模型引入额外的分布外检测模块，还是通过与人类价值观的契合来进行调整，都是具有重要研究价值的课题。

参考文献

[1] NI J, YOUNG T, PANDELEA V, et al. Recent advances in deep learning based dialogue systems: A systematic survey[J]. Artificial intelligence review, 2023, 56(4): 3055-3155.

[2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[3] KIM Y. Convolutional Neural Networks for Sentence Classification[C/OL]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1746-1751. https://aclanthology.org/D14-1181. DOI: 10.3115/v1/D14-1181.

[4] YOGATAMA D, DYER C, LING W, et al. Generative and discriminative text classification with recurrent neural networks[J]. ArXiv preprint arXiv:1703.01898, 2017.

[5] WU C S, HOI S C, SOCHER R, et al. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 917-929. DOI: 10.18653/v1/2020.emnlp-main.66.

[6] ZHOU Z H. Open-environment machine learning[J]. National Science Review, 2022, 9(8): nwac123.

[7] WU C S, MADOTTO A, HOSSEINI-ASL E, et al. Transferable multi-domain state generator for task-oriented dialogue systems[J]. ArXiv preprint arXiv:1905.08743, 2019.

[8] GANGADHARAIAH R, NARAYANASWAMY B. Recursive template-based frame generation for task oriented dialog[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2059-2064.

[9] OUYANG Y, CHEN M, DAI X, et al. Dialogue state tracking with explicit slot connection modeling[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 34-40.

[10] CHEN J, ZHANG R, MAO Y, et al. Parallel interactive networks for multi-domain dialogue state generation[J]. ArXiv preprint arXiv:2009.07616, 2020.

[11] GORDON-HALL G, GORINSKI P J, COHEN S B. Learning dialog policies from weak demonstrations[J]. ArXiv preprint arXiv:2004.11054, 2020.

[12] HUANG X, QI J, SUN Y, et al. Semi-supervised dialogue policy learning via stochastic reward estimation[J]. ArXiv preprint arXiv:2005.04379, 2020.

[13] LE H, SAHOO D, LIU C, et al. UniConv: A unified conversational neural architecture for multi-domain task-oriented dialogues[J]. ArXiv preprint arXiv:2004.14307, 2020.

[14] XU J, WANG H, NIU Z Y, et al. Conversational graph grounded policy learning for open-domain conversation generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 1835-1845.

[15] BAHETI A, RITTER A, SMALL K. Fluent response generation for conversational question answering[J]. ArXiv preprint arXiv:2005.10464, 2020.

[16] GOLOVANOV S, KURBANOV R, NIKOLENKO S, et al. Large-scale transfer learning for natural language generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 6053-6058.

[17] LI Y, YAO K, QIN L, et al. Slot-consistent NLG for task-oriented dialogue systems with iterative rectification network[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 97-106.

[18] BALAKRISHNAN A, RAO J, UPASANI K, et al. Constrained decoding for neural NLG from compositional representations in task-oriented dialogue[J]. ArXiv preprint arXiv:1906.07220, 2019.

[19] CHEN Q, ZHUO Z, WANG W. Bert for joint intent classification and slot filling[J]. ArXiv preprint arXiv:1902.10909, 2019.

[20] SCHUURMANS J, FRASINCAR F. Intent classification for dialogue utterances[J]. IEEE Intelligent Systems, 2019, 35(1): 82-88.

[21] WILSON T, WIEBE J, HOFFMANN P. Recognizing contextual polarity in phrase-level sentiment analysis[C]//Proceedings of human language technology conference and conference on empirical methods in natural language processing. 2005: 347-354.

[22] WILSON T, WIEBE J, HOFFMANN P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis[J]. Computational linguistics, 2009, 35(3): 399-433.

[23] SEBASTIANI F, ESULI A. Sentiwordnet: A publicly available lexical resource for opinion mining[C]//Proceedings of the 5th international conference on language resources and evaluation. 2006: 417-422.

[24] DONG Z, DONG Q, HAO C. Hownet and its computation of meaning[C]// Coling 2010: Demonstrations. 2010: 53-56.

[25] LEWIS D D. Naive (Bayes) at forty: The independence assumption in information retrieval[C]//European conference on machine learning. 1998: 4-15.

[26] NIGAM K, LAFFERTY J, MCCALLUM A. Using maximum entropy for text classification[C]//IJCAI-99 workshop on machine learning for information filtering: vol. 1: 1. 1999: 61-67.

[27] CORTES C, VAPNIK V. Support-vector networks[J]. Machine learning, 1995, 20: 273-297.

[28] SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1631-1642.

[29] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016: 1480-1489.

[30] HAKKANI-TÜR D, TÜR G, CELIKYILMAZ A, et al. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm.[C]//Interspeech. 2016: 715-719.

[31] CHEN Y N, HAKKANI-TÜR D, TÜR G, et al. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding.[C]//Interspeech. 2016: 3245-3249.

[32] ZHANG X, WANG H. A joint model of intent determination and slot filling for spoken language understanding.[C]//IJCAI: vol. 16. 2016: 2993-2999.

[33] LAMPERT C H, NICKISCH H, HARMELING S. Learning to detect unseen object classes by between-class attribute transfer[C]//2009 IEEE conference on computer vision and pattern recognition. 2009: 951-958.

[34] FU Y, HOSPEDALES T M, XIANG T, et al. Transductive multi-view zero-shot learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(11): 2332-2345.

[35] SUN P F, OUYANG Y W, SONG D J, et al. Self-Supervised Task Augmentation for Few-Shot Intent Detection[J]. Journal of Computer Science and Technology, 2022, 37(3): 527-538.

[36] KOCH G, ZEMEL R, SALAKHUTDINOV R, et al. Siamese neural networks for one-shot image recognition[C]//ICML deep learning workshop: vol. 2: 1.

[37] SARZYNSKA-WAWER J, WAWER A, PAWLAK A, et al. Detecting formal thought disorder by deep contextualized word representations[J]. Psychiatry Research, 2021, 304: 114135.

[38] ZHANG Y, ZHANG H, ZHAN L M, et al. New intent discovery with pre-training and contrastive learning[J]. ArXiv preprint arXiv:2205.12914, 2022.

[39] LIN T E, XU H, ZHANG H. Discovering new intents via constrained deep adaptive clustering with cluster refinement[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 34: 05. 2020: 8360-8367.

[40] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the national academy of sciences, 2017, 114(13): 3521-3526.

[41] LOPEZ-PAZ D, RANZATO M. Gradient episodic memory for continual learning[J]. Advances in neural information processing systems, 2017, 30.

[42] MADOTTO A, LIN Z, ZHOU Z, et al. Continual learning in task-oriented dialogue systems[J]. ArXiv preprint arXiv:2012.15504, 2020.

[43] ARORA U, HUANG W, HE H. Types of Out-of-Distribution Texts and How to Detect Them[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 10687-10701. DOI: 10.18653/v1/2021.emnlp-main.835.

[44] SHEN Z, LIU J, HE Y, et al. Towards out-of-distribution generalization: A survey[J]. ArXiv preprint arXiv:2108.13624, 2021.

[45] ZHANG X, HE Y, WANG T, et al. NICO Challenge: Out-of-Distribution Generalization for Image Recognition Challenges[C]//Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI. 2023: 433-450.

[46] ZHOU X, LIN Y, PI R, et al. Model agnostic sample reweighting for out-of-distribution learning[C]//International Conference on Machine Learning. 2022: 27203-27221.

[47] YE J, OUYANG Y, WU Z, et al. Out-of-Distribution Generalization Challenge in Dialog State Tracking[C]//NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications.

[48] OUYANG Y, YE J, CHEN Y, et al. Energy-based Unknown Intent Detection with Data Manipulation[Z]. 2021. arXiv: 2107.12542.

[49] OUYANG Y, WU Z, DAI X, et al. Towards Multi-label Unknown Intent Detection[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 626-635.

[50] LIN T E, XU H. Deep Unknown Intent Detection with Margin Loss[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5491-5496. DOI: 10.18653/v1/P19-1548.

[51] REN J, LIU P J, FERTIG E, et al. Likelihood ratios for out-of-distribution detection[C]//Advances in Neural Information Processing Systems. 2019: 14680-14691.

[52] LIU J, LIN Z, PADHY S, et al. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness[J]. Advances in Neural Information Processing Systems, 2020, 33: 7498-7512.

[53] YANG J, ZHOU K, LI Y, et al. Generalized out-of-distribution detection: A survey[J]. ArXiv preprint arXiv:2110.11334, 2021.

[54] NALISNICK E, MATSUKAWA A, TEH Y W, et al. Do deep generative models know what they don't know?[J]. ArXiv preprint arXiv:1810.09136, 2018.

[55] SERRÀ J, ÁLVAREZ D, GÓMEZ V, et al. Input Complexity and Out-of-distribution Detection with Likelihood-based Generative Models[C]//International Conference on Learning Representations. 2020.

[56] HENDRYCKS D, GIMPEL K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks[J]. Proceedings of International Conference on Learning Representations, 2017.

[57] LEE K, LEE K, LEE H, et al. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks[G]//BENGIO S, WALLACH H, LAROCHELLE H, et al. Advances in Neural Information Processing Systems 31. Curran Associates, Inc., 2018: 7167-7177.

[58] LIU W, WANG X, OWENS J, et al. Energy-based Out-of-distribution Detection[J]. Advances in Neural Information Processing Systems, 2020, 33.

[59] VAZE S, HAN K, VEDALDI A, et al. Open-Set Recognition: A Good Closed-Set Classifier is All You Need[C]//International Conference on Learning Representations. 2022.

[60] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. ArXiv preprint arXiv:2010.11929, 2020.

[61] LEE K, LEE H, LEE K, et al. Training confidence-calibrated classifiers for detecting out-of-distribution samples[J]. ArXiv preprint arXiv:1711.09325, 2017.

[62] KAMOI R, KOBAYASHI K. Why is the mahalanobis distance effective for anomaly detection?[J]. ArXiv preprint arXiv:2003.00402, 2020.

[63] WINKENS J, BUNEL R, ROY A G, et al. Contrastive training for improved out-of-distribution detection[J]. ArXiv preprint arXiv:2007.05566, 2020.

[64] LECUN Y, CHOPRA S, HADSELL R, et al. A tutorial on energy-based learning[J]. Predicting structured data, 2006, 1(0).

[65] RADFORD A, WU J, CHILD R, et al. Language Models are Unsupervised Multitask Learners[J]., 2019.

[66] BROWN T, MANN B, RYDER N, et al. Language Models are Few-Shot Learners[C]//LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems: vol. 33. Curran Associates, Inc., 2020: 1877-1901.

[67] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. ArXiv preprint arXiv:1312.6114, 2013.

[68] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[69] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[70] YANG J, WANG P, ZOU D, et al. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection[C]//Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track. 2022.

[71] RUFF L, KAUFFMANN J R, VANDERMEULEN R A, et al. A unifying review of deep and shallow anomaly detection[J]. ArXiv preprint arXiv:2009.11732, 2020.

[72] PANG G, SHEN C, CAO L, et al. Deep learning for anomaly detection: A review[J]. ACM computing surveys (CSUR), 2021, 54(2): 1-38.

[73] SCHEIRER W J, de REZENDE ROCHA A, SAPKOTA A, et al. Toward open set recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(7): 1757-1772.

[74] BASU S, MECKESHEIMER M. Automatic outlier detection for time series: an application to sensor data[J]. Knowledge and Information Systems, 2007, 11: 137-154.

[75] WANG J, LAN C, LIU C, et al. Generalizing to Unseen Domains: A Survey on Domain Generalization[C]//ZHOU Z H. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. International Joint Conferences on Artificial Intelligence Organization, 2021: 4627-4635.

[76] WANG M, DENG W. Deep visual domain adaptation: A survey[J]. Neurocomputing, 2018, 312: 135-153.

[77] DAVIS J, GOADRICH M. The relationship between Precision-Recall and ROC curves[C]//Proceedings of the 23rd international conference on Machine learning. 2006: 233-240.

[78] MANNING C D, MANNING C D, SCHÜTZE H. Foundations of statistical natural language processing[M]. MIT press, 1999.

[79] ZHOU Z H. Machine learning[M]. Springer Nature, 2021.

[80] 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 (EMNLP-IJCNLP). 2019: 1311-1316. DOI: 10.18653/v1/D19-1131.

[81] GURURANGAN S, SWAYAMDIPTA S, LEVY O, et al. Annotation Artifacts in Natural Language Inference Data[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 107-112 [2023-03-13].

[82] WILLIAMS A, NANGIA N, BOWMAN S R. A broad-coverage challenge corpus for sentence understanding through inference[J]. ArXiv preprint arXiv:1704.05426, 2017.

[83] WANG Z, CULOTTA A. Identifying Spurious Correlations for Robust Text Classification[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 3431-3440. DOI: 10.18653/v1/2020.findings-emnlp.308.

[84] SCHUSTER T, SHAH D, YEO Y J S, et al. Towards Debiasing Fact Verification Models[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3419-3425.

[85] ZHANG Y, BALDRIDGE J, HE L. PAWS: Paraphrase Adversaries from Word Scrambling[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). 2019: 1298-1308.

[86] ZHENG Y, CHEN G, HUANG M. Out-of-domain Detection for Natural Language Understanding in Dialog Systems[J]. ArXiv preprint arXiv:1909.03862, 2019.

[87] SHRIKUMAR A, GREENSIDE P, SHCHERBINA A, et al. Not just a black box: Learning important features through propagating activation differences[J]. ArXiv preprint arXiv:1605.01713, 2016.

[88] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. ArXiv preprint arXiv:1810.04805, 2018.

[89] CHEN D, YU Z. Gold: improving out-of-scope detection in dialogues using data augmentation[J]. ArXiv preprint arXiv:2109.03079, 2021.

[90] GANGADHARAIAH R, NARAYANASWAMY B. Joint multiple intent detection and slot labeling for goal-oriented dialog[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). 2019: 564-569.

[91] WANG R, DAI X, et al. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2022: 672-679.

[92] WANG R, LONG S, DAI X, et al. Meta-LMTC: meta-learning for large-scale multi-label text classification[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 8633-8646.

[93] WANG R, RIDLEY R, QU W, et al. A novel reasoning mechanism for multi-label text classification[J]. Information Processing & Management, 2021, 58(2): 102441.

[94] BOUTELL M R, LUO J, SHEN X, et al. Learning multi-label scene classification[J]. Pattern recognition, 2004, 37(9): 1757-1771.

[95] READ J, PFAHRINGER B, HOLMES G, et al. Classifier chains for multi-label classification[J]. Machine learning, 2011, 85(3): 333-359.

[96] YANG P, SUN X, LI W, et al. SGM: Sequence Generation Model for Multi-label Classification[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 3915-3926.

[97] QIN L, XU X, CHE W, et al. AGIF: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 1807-1816. DOI: 10.18653/v1/2020.findings-emnlp.163.

[98] HOU Y, LAI Y, WU Y, et al. Few-shot Learning for Multi-label Intent Detection[C]//Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021. AAAI Press, 2021: 13036-13044.

[99] YAN G, FAN L, LI Q, et al. Unknown Intent Detection Using Gaussian Mixture Model with an Application to Zero-shot Intent Classification[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 1050-1060. DOI: 10.18653/v1/2020.acl-main.99.

[100] MURPHY K P. Probabilistic machine learning: an introduction[M]. MIT press, 2022.

[101] SHU L, XU H, LIU B. DOC: Deep Open Classification of Text Documents[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2911-2916. DOI: 10.18653/v1/D17-1314.

[102] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. ArXiv preprint arXiv:1412.3555, 2014.

[103] MULLENBACH J, WIEGREFFE S, DUKE J, et al. Explainable Prediction of Medical Codes from Clinical Text[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 1101-1111. DOI: 10.18653/v1/N18-1100.

[104] HAN T, LIU X, TAKANOBU R, et al. MultiWOZ 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation[J]. ArXiv preprint arXiv:2010.05594, 2020.

[105] COUCKE A, SAADE A, BALL A, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces[J]. ArXiv preprint arXiv:1805.10190, 2018.

[106] GANGAL V, ARORA A, EINOLGHOZATI A, et al. Likelihood Ratios and Generative Classifiers for Unsupervised Out-of-Domain Detection in Task Ori-

ented Dialog[C]//The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020. AAAI Press, 2020: 7764-7771.

[107] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density-based local outliers[C]//Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 2000: 93-104.

[108] PANG T, XU K, DONG Y, et al. Rethinking Softmax Cross-Entropy Loss for Adversarial Robustness[C]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

[109] SCHIRRMEISTER R, ZHOU Y, BALL T, et al. Understanding Anomaly Detection with Deep Invertible Networks through Hierarchies of Distributions and Features[C]//LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems: vol. 33. Curran Associates, Inc., 2020: 21038-21049.

[110] GANGAL V, ARORA A, EINOLGHOZATI A, et al. Likelihood Ratios and Generative Classifiers for Unsupervised Out-of-Domain Detection In Task Oriented Dialog[J]. ArXiv preprint arXiv:1912.12800, 2019.

[111] KHANDELWAL U, LEVY O, JURAFSKY D, et al. Generalization through memorization: Nearest neighbor language models[J]. ArXiv preprint arXiv:1911.00172, 2019.

[112] KHANDELWAL U, FAN A, JURAFSKY D, et al. Nearest neighbor machine translation[J]. ArXiv preprint arXiv:2010.00710, 2020.

[113] GU J, WANG Y, CHO K, et al. Search Engine Guided Non-Parametric Neural Machine Translation[J]. ArXiv preprint arXiv:1705.07267, 2017.

[114] BORGEAUD S, MENSCH A, HOFFMANN J, et al. Improving language models by retrieving from trillions of tokens[C]//International conference on machine learning. 2022: 2206-2240.

[115] LI J, CHENG S, SUN Z, et al. Better Datastore, Better Translation: Generating Datastores from Pre-Trained Models for Nearest Neural Machine Translation[J]. ArXiv preprint arXiv:2212.08822, 2022.

[116] ZHU W, HUANG S, LV Y, et al. What Knowledge Is Needed? Towards Explainable Memory for kNN-MT Domain Adaptation[J]. ArXiv preprint arXiv:2211.04052, 2022.

[117] ZHENG X, ZHANG Z, GUO J, et al. Adaptive nearest neighbor machine translation[J]. ArXiv preprint arXiv:2105.13022, 2021.

[118] JIANG Q, WANG M, CAO J, et al. Learning kernel-smoothed machine translation with retrieved examples[J]. ArXiv preprint arXiv:2109.09991, 2021.

[119] WANG D, WEI H, ZHANG Z, et al. Non-parametric online learning from human feedback for neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 36: 10. 2022: 11431-11439.

[120] ZHENG X, ZHANG Z, HUANG S, et al. Non-parametric unsupervised domain adaptation for neural machine translation[J]. ArXiv preprint arXiv:2109.06604, 2021.

[121] KASSNER N, SCHÜTZE H. BERT-kNN: Adding a kNN search component to pretrained language models for better QA[J]. ArXiv preprint arXiv:2005.00766, 2020.

[122] HE J, NEUBIG G, BERG-KIRKPATRICK T. Efficient nearest neighbor language models[J]. ArXiv preprint arXiv:2109.04212, 2021.

[123] SUNDERMEYER M, SCHLÜTER R, NEY H. LSTM neural networks for language modeling[C]//Thirteenth annual conference of the international speech communication association. 2012.

[124] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[125] PODOLSKIY A, LIPIN D, BOUT A, et al. Revisiting Mahalanobis Distance for Transformer-Based Out-of-Domain Detection[C]//Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021. AAAI Press, 2021: 13675-13682.

[126] LANG H, ZHENG Y, SUN J, et al. Estimating Soft Labels for Out-of-Domain Intent Detection[J]. ArXiv preprint arXiv:2211.05561, 2022.

[127] BLACK S, BIDERMAN S, HALLAHAN E, et al. Gpt-neox-20b: An open-source autoregressive language model[J]. ArXiv preprint arXiv:2204.06745, 2022.

[128] ZHANG S, ROLLER S, GOYAL N, et al. Opt: Open pre-trained transformer language models[J]. ArXiv preprint arXiv:2205.01068, 2022.

[129] ZENG A, LIU X, DU Z, et al. Glm-130b: An open bilingual pre-trained model[J]. ArXiv preprint arXiv:2210.02414, 2022.

[130] SCAO T L, FAN A, AKIKI C, et al. Bloom: A 176b-parameter open-access multilingual language model[J]. ArXiv preprint arXiv:2211.05100, 2022.

[131] MUENNIGHOFF N, WANG T, SUTAWIKA L, et al. Crosslingual generalization through multitask finetuning[J]. ArXiv preprint arXiv:2211.01786, 2022.

[132] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[J]. ArXiv preprint arXiv:2303.18223, 2023.

[133] GENG B, YUAN F, XU Q, et al. Continual learning for task-oriented dialogue system with iterative network pruning, expanding and masking[J]. ArXiv preprint arXiv:2107.08173, 2021.

[134] 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 (Volume 1: Long Papers). 2021: 4582-4597. DOI: 10.18653/v1/2021.acl-long.353.

[135] SUN Y, MING Y, ZHU X, et al. Out-of-Distribution Detection with Deep Nearest Neighbors[C]//International Conference on Machine Learning. 2022.

[136] XU K, REN T, ZHANG S, et al. Unsupervised Out-of-Domain Detection via Pre-trained Transformers[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 1052-1061. DOI: 10.18653/v1/2021.acl-long.85.

[137] GANGAL V, ABHINAV A, ARASH E, et al. Likelihood Ratios and Generative Classifiers for Unsupervised Out-of-Domain Detection in Task Oriented Dialog.[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 7764-7771.

[138] SALEHI M, MIRZAEI H, HENDRYCKS D, et al. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges[J]. ArXiv preprint arXiv:2110.14051, 2021.

[139] LIU P, YUAN W, FU J, et al. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing[J]. ACM Comput. Surv., 2022. DOI: 10.1145/3560815.

[140] QIU X, SUN T, XU Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China Technological Sciences, 2020, 63(10): 1872-1897.

[141] LIU X, HUANG H, SHI G, et al. Dynamic Prefix-Tuning for Generative Template-based Event Extraction[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 5216-5228. DOI: 10.18653/v1/2022.acl-long.358.

[142] ZHAO L, ZHENG F, ZENG W, et al. Domain-Oriented Prefix-Tuning: Towards Efficient and Generalizable Fine-tuning for Zero-Shot Dialogue Summarization[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022: 4848-4862. DOI: 10.18653/v1/2022.naacl-main.357.

[143] CHEN Y, LIU Y, XU R, et al. UniSumm: Unified Few-shot Summarization with Multi-Task Pre-Training and Prefix-Tuning[J]. ArXiv preprint arXiv:2211.09783, 2022.

[144] MA Y, NGUYEN T H, MA B. CPT: Cross-Modal Prefix-Tuning for Speech-To-Text Translation[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022: 6217-6221. DOI: 10.1109/ICASSP43922.2022.9746935.

[145] LIU X, JI K, FU Y, et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks[J]. CoRR, 2021, abs/2110.07602. arXiv: 2110.07602.

[146] YANG Z, LIU Y. On Robust Prefix-Tuning for Text Classification[C]//International Conference on Learning Representations. 2022.

[147] BALAKRISHNAN S, FANG Y, ZHU X. Exploring Robustness of Prefix Tuning in Noisy Data: A Case Study in Financial Sentiment Analysis[J]. ArXiv preprint arXiv:2211.05584, 2022.

[148] WOLF T, DEBUT L, SANH V, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing[J]. ArXiv, 2019, abs/1910.03771.

[149] DING N, HU S, ZHAO W, et al. OpenPrompt: An Open-source Framework for Prompt-learning[J]. ArXiv preprint arXiv:2111.01998, 2021.

[150] PENNINGTON J, SOCHER R, MANNING C. GloVe: Global Vectors for Word Representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543. DOI: 10.3115/v1/D14-1162.

[151] SCHICK T, SCHÜTZE H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 255-269. DOI: 10.18653/v1/2021.eacl-main.20.

[152] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. 2021: 8748-8763.

[153] HUANG S, DONG L, WANG W, et al. Language is not all you need: Aligning perception with language models[J]. ArXiv preprint arXiv:2302.14045, 2023.

[154] TSIMPOUKELLI M, MENICK J L, CABI S, et al. Multimodal few-shot learning with frozen language models[J]. Advances in Neural Information Processing Systems, 2021, 34: 200-212.

[155] LI J, LI D, SAVARESE S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[J]. ArXiv preprint arXiv:2301.12597, 2023.

[156] SU W, ZHU X, CAO Y, et al. Vl-bert: Pre-training of generic visual-linguistic representations[J]. ArXiv preprint arXiv:1908.08530, 2019.

致　　谢

弹指间，已经度过了五年的直博时光，回望这段时光中的点点滴滴，我有太多需要感谢的人。

首先，衷心感谢我的导师戴新宇老师。科研上，戴老师给予了我极大的帮助，读博初期日常的讨论，周报的点评，论文的修改，这些全方位的指导让我学会了独立研究者应具有的素质。生活中，戴老师也经常为我排忧解难，生病时的关心，受挫时的鼓励，都让我倍感温暖。我仍能清楚记得戴老师给我发短信，通知我研究生录取的那个夏日傍晚，再次感激戴老师的知遇之恩。此外，也要感谢NJUNLP 组的陈家骏老师、黄书剑老师、张建兵老师、吴震老师、何亮老师，还有滑铁卢大学的宗石老师，每次开大组会时，都能感受到你们对待科研问题的谦虚谨慎，你们一直是我科研道路和人生道路上的指路明灯。

其次，感谢字节跳动人工智能实验室的李航老师和弗吉尼亚大学的王宏宁老师，感谢你们在我实习期间对我耐心和全面的指导。感谢字节跳动的黄浩然、徐爽等同事，与你们的每一次讨论都让我收获满满。感谢上海疫情期间，照顾我生活的朋友们，尤其是我的发小王烽先，是你们帮我度过了人生中最黑暗的岁月。

感谢研究组的孙鹏飞博士、鲍宇博士、郑在翔博士、汪然博士、周浩博士、陈华栋博士、李家欢、邹威、琚江舟、Robert Ridley、赵飞、马征、龙思宇、朱文昊、陶堃、宁天昊、邱丰羽、范志方、成臻、谭春阳、杨凯嘉、杨惠云、许喆、张文明、陈陌信、孟凡杰、崔渠、周志浩、陈昱、宋定杰、高源、曹永昌、王仁杰、吕云哲、汤远航、叶家升、刘莉以及其他同学，这些年来，正是因为你们的陪伴，使得我在读博期间不再孤单。

感谢我的家人对我求学路上，一路走来的无条件支持，读书这么多年，一直没能为家里做点什么，但你们未曾抱怨，并一直做我最坚实的后盾。感谢我的女友王舒颖女士对我的陪伴，每天能听到你的声音，我觉得很幸福。希望你们永远健康且快乐。

博士毕业不是终点，我深知自己还有太多需要改进和提升的地方，希望自己能够更加勇敢，在人生马拉松中不断奋进，诚挚地面对一切，无愧于帮助我的人们。

最后我想以香港岭南大学李连江老师曾说过的一段话作为结尾：走在路上总是辛苦的。但是，有路可走就是幸福，走得动路是更大的幸福。

科研成果与学术活动

科研成果

1. **Yawen Ouyang**, Yongchang Cao, Yuan Gao, Zhen Wu, Jianbing Zhang and Xinyu Dai. "On Prefix-tuning for Lightweight Out-of-distribution Detection," in *Proceedings of the 61th annual meeting of the association for computational linguistics.* (**ACL 2023, CCF A 类会议**)
2. **Yawen Ouyang**, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. "Towards Multi-label Unknown Intent Detection," in *Proceedings of the 29th International Conference on Computational Linguistics.* (**COLING 2022, CCF B 类会议**)
3. **Yawen Ouyang**, Jiasheng Ye, Yu Chen, Xinyu Dai, Shujian Huang, and Jiajun Chen. "Energy-based Unknown Intent Detection with Data Manipulation," in *Proceedings of the 59th annual meeting of the association for computational linguistics.* (**Findings of ACL 2021**)
4. **Yawen Ouyang**, Moxin Chen, Xinyu Dai, Yinggong Zhao, Shujian Huang, and Jiajun Chen. "Dialogue state tracking with explicit slot connection modeling," in *Proceedings of the 58th annual meeting of the association for computational linguistics, short paper.* (**ACL 2020, CCF A 类会议**)
5. **Yawen Ouyang**, Yuan Gao, Shi Zong, Yu Bao, and Xinyu Dai. "Pobe, a generative-based out-of-distribution text detection method" in *Journal of Software.* (**软件学报, CCF T1 类期刊**)
6. Pengfei Sun, **Yawen Ouyang**, Wenming Zhang, and Xinyu Dai. "MEDA: Meta-Learning with Data Augmentation for Few-Shot Text Classification," in *IJCAI.* (**IJCAI 2021, CCF A 类会议**)
7. Pengfei Sun, **Yawen Ouyang**, Dingjie Song, and Xinyu Dai. "Self-Supervised Task Augmentation for Few-Shot Intent Detection," in *Journal of Computer Science and*

Technology, 37(3), 527-538. (**JCST, CCF B 类期刊**)

8. Yuanhang Tang, **Yawen Ouyang**, Zhen Wu, Baohua Zhang, Jiaying Zhang and Xinyu Dai. "IDOS: A Unified Debiasing Framework via Word Shuffling," in *NLPCC 2023.* (**NLPCC 2023, CCF C 类会议**)
9. Pengfei Sun, **Yawen Ouyang**, Xinyu Dai, and Wenming Zhang. "Similarity Learning with Implicit-network and Explicit-network for Zero-shot Intent Detection," in *SCIENTIA SINICA Informationis.* (**中国科学: 信息科学, CCF T1 类期刊**)
10. Jiasheng Ye, **Yawen Ouyang**, Zhen Wu, and Xinyu Dai. "Out-of-Distribution Generalization Challenge in Dialog State Tracking," in *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications.*
11. Pengfei Sun, Dingjie Song, **Yawen Ouyang**, Zhen Wu, and Xinyu Dai. "Episode-based Prompt Learning for Any-shot Intent Detection," in *NLPCC 2023.* (**NLPCC 2023, CCF C 类会议**)

担任审稿人

- CCF A 类会议: NeurIPS, ICML, AAAI
- CCF B 类会议: EMNLP, NAACL
- CCF C 类会议: NLPCC, IJCNN
- 其他: ICLR

学位论文出版授权书

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》（以下简称“章程”），愿意将本人的学位论文提交“中国学术期刊（光盘版）电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名：_____

_____年____月____日

论文题名	面向任务型对话系统的分布外检测方法研究				
研究生学号	DZ1833020	所在院系	计算机科学与技术	学位年度	2023
论文级别	☐ 学术学位硕士　☐ 专业学位硕士 ☑ 学术学位博士　☐ 专业学位博士				
作者 Email	ouyangyw@smail.nju.edu.cn				
导师姓名	戴新宇				

论文涉密情况：

☐ 不保密

☐ 保密，保密期（_____年____月____日至_____年____月____日）