附件1：

# “华为杯”第四届中国研究生人工智能创新大赛赛题

**一、大赛赛题**

（一）技术创新：开放题

（二）应用创意：开放题

（三）企业赛题：华为赛题共4个，均为华为技术有限公司从实际需求出发拟定的与智能技术密切相关的题目。属于此种选题方式的作品还可以参与华为专项奖的评奖。

**1.华为赛题一**

1.1 题目名称：微服务架构应用的故障演练数字化

1.2 题目描述

随着数字化程度越来越高，分布式系统、云计算等新技术的不断发展，运维在企业中的重要性越来越高，用户和企业从质量、效率、性能、安全、成本等多维度出发关注运维。且随着企业IT系统的规模扩大、复杂度不断提高、监控数据日益增长，各类故障层出不穷，保证系统高效可靠运转的难度激增，运维需要与人工智能的能力结合，基于企业IT运维过程中产生的告警事件、黄金指标、日志、调用链等数据，利用机器学习/深度学习/数据挖掘相关技术支撑异常检测、根因定位和容量规划等场景，助力企业应用稳定、高效地运行，解决传统运维难以解决的问题，提升运维效率，为企业减少因业务不可用所带来的直接损失。

现代IT系统是一个复杂系统，随着应用现代化的推进，越来越多传统单体应用转为微服务架构，微服务数量巨大，关系复杂，迭代变化快，传统运维越来越困难。该方向旨在通过各类异构多源数据的采集、感知与处理，使用多维时序分析、神经网络、机器学习、复杂系统控制论、关系图谱等技术，完成对微服务系统的异常检测、日志分析、调用链分析，和故障定位来帮助IT系统提前发现和预知风险，维持系统稳定与安全运行，在系统发生故障时，帮助运维人员快速定位及解决故障，从而保障系统的高可靠、高可用性。

主要挑战:

（1）真实场景里的故障样本与正常样本相比数据量较少，有数据不平衡的问题，很多数据没有标签, 也有数据噪声大，信噪比低的问题，对AI模型的开发不友好。

（2）IT系统的数据形态非常多样, 有结构化的监控指标数据，有半结构化的调用链数据，及非结构化的告警和日志类数据，鲁邦的AIOps模型需融合多源异构数据，从不同数据源提取特征构建。

IT应用在上线前需要经过测试环节，上线后也有例行的故障演练。在测试和故障演练的过程中，通过有效的挖掘导致应用故障的失效模式，记录应用的各类数据，可以形成AI友好的完整数据集支撑后续的异常检测与根因定位模型开发，参赛者需要设计失效模式挖掘方案让故障演练更好地数字化。

1.3 具体要求

（1）参赛者需要部署微服务Microservices-demo[1]及配套监控系统(包括但不限于[2]、[3])采集指标数据；

（2）参赛者需要考虑用户访问模型，模拟不同的用户访问模式，利用故障注入工具(包括但不限于[4])使微服务应用出现故障，挖掘微服务应用的失效模式，常见的失效模式包括：前端负荷激增(业务量异常)、用户打开页面卡顿， 用户打不开页面等失效模式，它们可能是由数据库服务不可用(连接数达到上限, 数据库表被删除,数据库表过大)、底层资源容量不足等（CPU、内存利用率…）、网络阻塞等引起的；

（3）参赛者生成故障数据集，并通过案例展示该生成数据集可支撑异常检测与根因定位算法[5]-[9]的开发与验证。

（4）参赛队伍最终提交故障数据集，模型代码，方案文档与演示视频(可选),方案文档需完整描述故障演练数字化方案与AI模型训练与验证过程。

a.生成的故障数据需要有多样性，至少覆盖3种组件与3类典型故障；

b.异常检测与根因定位算法的评测指标可采用业界常见的评估方法，异常检测算法可采用F1-Score作为评价指标（参赛者需根据应用的访问延时与错误率等黄金指标定义异常,为数据打上标签后再验证算法准确率），根因定位算法可采用PR@k（Precision at top k, top k个结果中包含真实根因的概率）与MAP（Mean Average Precision）作为评价指标,同等准确率的情况下,算法的推理时间与模型大小将作为参考评价指标。

c.人工评委审核成绩从4个方面考核：结果复现评估，方案可行性分析，方案创新性评估，方案通用性评估；评委人员：潘乐萌，张晓，白志刚，郭毅成。

1.4华为赛题一咨询专家及联系邮箱：

微服务架构应用的故障演练数字化–潘乐萌- panlemeng@huawei.com

1.5参考数据集

AIOps-Challenge-2020-Data:<https://github.com/NetManAIOps/AIOps-Challenge-2020-Data>

MicroRCA Sample Data: <https://github.com/elastisys/MicroRCA/tree/master/data>

1.6参考文献

MMicroservices-demo

[1] https://github.com/microservices-demo/microservices-demo

Prometheus - Monitoring system & time series database

[2]https://prometheus.io

Grafana: The open observability platform

[3]https://grafana.com/

Locust - A modern load testing framework

[4]https://locust.io

Anomaly Detection

[5] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey[J]. ACM computing surveys (CSUR), 2009, 41(3): 1-58.

[6] Chalapathy R, Chawla S. Deep learning for anomaly detection: A survey[J]. arXiv preprint arXiv:1901.03407, 2019.

Root Cause Analysis

[7] Wu L, Tordsson J, Elmroth E, et al. Microrca: Root cause localization of performance issues in microservices[C]//NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium. IEEE, 2020: 1-9.

[8] Lin J J, Chen P, Zheng Z. Microscope: Pinpoint performance issues with causal graphs in micro-service environments[C]//International Conference on Service-Oriented Computing. Springer, Cham, 2018: 3-20.

[9] Chen M Y, Kiciman E, Fratkin E, et al. Pinpoint: Problem determination in large, dynamic internet services[C]//Proceedings International Conference on Dependable Systems and Networks. IEEE, 2002: 595-604.

**2.华为赛题二**

2.1 题目名称:面向口语的NL2SQL语义解析算法

2.2 题目描述

语义解析技术是自然语言与企业数据之间的桥梁，为企业级数据查询的易用性、快捷性发挥着重要作用，为提高企业的数字化转型有着重大意义。当前的技术方案并非直接将自然语言转为SQL形式，而是识别出Query中的各个SQL条件成分、条件值等要素，以生成伪SQL形式。然而，自然语言的多样性限制了NL2SQL场景的能力，在实际应用过程中会存在如下问题：

（1）**Query口语化（重点）**。用户日常查表过程中，口语化成分较多，指代复杂，歧义性较高，如：90后员工有多少人、去年部门走了多少员工等，难以用程序规则覆盖完整，存在“长尾”问题。

（2）**小样本训练能力**。企业级的表格都是大宽表形式，按照主流NL2SQL或语义解析做法，需要进行标注的数据量必须涵盖所有字段，标注任务繁重，且算法能力后期难以迁移至其他企业业务场景；目前暂不涉及多表查询的需求，但单表字段较复杂，有近义词的字段，有字符不统一，有长文本字段等。

2.3 具体要求

（1）所提方法，必须具有“可迁移性”，不可只能解决某一个业务领域问题。

（2）可利用外部知识（如：公开数据集、自有数据知识、适当程序正则等）增强模型算法，并最终在业务测试集上进行测试，证明方法的有效性。

（3）在保证模型“可迁移性”条件下，尽量保证模型精度不下降，最终以验证集评测结果为首要标准，以方案可行性为辅助评测标准。

（4）由于此课题主要是针对SQL表字段及字段值，进行Query的语义成分解析，可能涉及实体识别、句子成分解析等技术，同时也鼓励新技术新idea创新。以SQL单表为例，单表字段“时间；产品类别；销售状态”，示例Query“2020年有多少C类产品售出？”，需要解析成以下条件（单条query所有要素识别正确加1分）：

a.先识别此query的操作为SQL的COUNT操作，以“有多少”为依据，总共有六种操作“MAX\MIN\COUNT\AVG\SUM\NULL”，其中“NULL”为无关语料不需要做解析（other类问题）；

b.条件值“2020年”对应时间字段的条件值，“C类产品”对应产品类别的条件值，“售出”对应销售状态的条件值；

（5）最终打榜成绩（计60%分数）。通过线下脚本计算判分取Top10进行人工评委审核，若存在并列成绩情况，可多取候选者；赛委会提供1张宽表（约20-30个字段），以及1份已标注的Query数据集，赛委会保留1份待预测数据集作为评测，结果按照约定JSON格式提交，以正确率为评判标准（单条所有要素全部正确匹配才算得分，参见2.3部分解析），最终评委组通过线下自动化脚本判分，以邮件或官网形式公布最后的打榜成绩；

（6）人工评委审核成绩（计40%分数）。评审从4个方面考核：打榜结果复现评估，方案可行性分析，方案创新性评估，方案可迁移性评估；初审评委人员：夏海荣、周宇翔、孙非凡、蒋圭峰。

2.4 华为赛题—咨询专家及联系邮箱

面向口语的NL2SQL语义解析算法 –周宇翔 –zhouyuxiang4@huawei.com

**3.华为赛题三**

3.1 题目名称：智能问答中的澄清问的识别与生成

3.2 题目描述

由于自然语言的复杂性和现实世界里知识的多样性，智能问答系统在部分对话场景下需要向用户发起追问，以更加准确地理解用户的真实意图。这类问题通常称为“澄清问”，旨在引导用户对其意图进行明确和细化，从而消除对话中的语义歧义和知识歧义，提高智能问答系统的效果。因此，澄清问的检测、生成和答案理解，对于智能问答系统的构建具有重要意义。

具体地，该任务通常涉及澄清检测、澄清问生成以及实体预测三个子任务：

1.澄清检测

澄清检测任务需要考虑对话上下文，并结合问题的具体领域和所提供的知识库信息，对“当前语境下系统是否需要进行澄清问”进行判别。该任务通常可以视为二分类任务。

通常而言，该任务的输入是从0到t时刻的用户及系统历史对话<U0, M0, U1, M1, ..., Ut>，其中Uk为第k轮用户会话，Mk为第k轮系统会话；输出为系统对“当前是否需要进行澄清问”的预测结果，若需要进行澄清问则输出标签“1”，否则输出为“0”。

例如，假设知识库中包含歌手“孙悦”和篮球运动员“孙悦”的相关知识。当用户询问“孙悦是哪一年出生的”时，由于无法直接判断用户询问的是哪一位孙悦的信息，因此需要进行澄清问；当用户询问“孙悦效力于哪个篮球队”时，结合知识信息的简单推理可以判断用户询问的是篮球运动员的信息，此时不需要进行澄清问。

2.澄清问生成

若澄清检测结果为真，则智能问答系统需要生成相应的澄清提问。该任务可以视为文本生成任务，其输入同样是是从0到t时刻的用户及系统历史对话<U0, M0, U1, M1, ..., Ut>，其中Uk为第k轮用户会话，Mk为第k轮系统会话；输出为系统的澄清提问Q。

3.实体预测

在进行澄清提问后，系统需要进一步对用户的回答进行理解。该任务的输入表示为<U0, M0, U1, M1, ..., Ut, Q, A>，其中Uk为第k轮用户会话，Mk为第k轮系统会话，Q为系统的澄清提问，A为用户的回答。模型需要提取出A中指向的是哪个实体。

3.3 具体要求

（1）根据所公布的数据集，参赛者可自由设计算法方案来比赛，但不可调用第三方外部接口。

（2）参赛选手需完成全部三个子任务。其中，每个子任务均包含single-turn和multi-turn两种类型的数据，选手应采用统一的模型处理这两类数据（即，单轮与多轮不分开评估）。

（3）赛题总分由竞赛得分（60%）和专家评分（40%）两部分组成。其中，竞赛得分部分根据选手提交的方案在测试集（包括公开测试集和非公开测试集）上的运行结果给出；专家评分由评委组对选手所提交的方案的新颖性、合理性等进行打分。因此，除结果文件外，参赛选手还需要提交模型代码（用于非公开测试集评估）和模型说明文件（用于报告模型方案以及模型在公开测试集上的结果）。

3.4 华为赛题—咨询专家及联系邮箱

智能问答中的澄清问的识别与生成 –张睿、王喆锋–zhangrui311@huawei.com、 wangzhefeng@huawei.com

3.5参考数据集

CLAQUA数据集 <https://github.com/msra-nlc/MSParS_V2.0>

3.6参考文献

1. X u J, Wang Y, Tang D, et al. Asking clarification questions in knowledge-based question answering[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: 1618-1629.

2.  Shao T, Cai F, Chen W, et al. Self-supervised clarification question generation for ambiguous multi-turn conversation[J]. Information Sciences, 2022, 587: 626-641.

**4.华为赛题四**

4.1题目名称

预训练中的数据选取

4.2题目描述

预训练大模型需要从海量数据中进行学习。对于NLP预训练来说，训练语料通常需要几十GB到几百GB，如此大规模的数据，每一个训练轮次都要消耗大量算力和时间，带来了极高的训练成本。一个已有的观察是，在海量的训练数据中往往包含一些重复或相似的数据，如果对这些数据进行精简，可以在不影响模型性能的前提下，有效地压缩数据体积。此外，由于数据来源多样，训练数据中通常也会包含一些噪声数据，这些数据对模型的训练增益不大，甚至可能带来负面影响。

本赛题考虑对海量文本数据进行选取，从中剔除重复、相似以及有害的数据，从而压缩数据体积，并且尽可能不影响NLP预训练模型的性能。

4.3具体要求

给定10GB英文互联网数据。要求参赛者设计数据选取算法，将该数据压缩至70%，基于压缩后数据训练英文bert-base模型，训练轮次限定为3轮。训练完成后，在公开数据集SST-2、MNLI、SQuAD1.1上进行测试，根据测试集上的性能进行初筛，选出前五名进入第二阶段测试。要求参赛者提交选取算法代码、数据压缩记录、训练过程记录、测试结果截图等相关证明材料，供主办方审核。

第二阶段要求前五名参赛队伍提交完整的数据选取代码，主办方使用另外的20GB英文互联网数据，基于参赛队伍的算法进行数据压缩，得到70%的压缩数据，训练英文bert-large模型，并在GLUE和SuperGLUE两个公开榜单上进行测试。根据测试性能得到最终排名。

赛题输入

10GB英文数据。

赛题输出

压缩至70%，尽可能保证预训练模型在下游任务的性能。

4.4华为赛题—咨询专家及联系邮箱

预训练中的数据选取 –史佳欣— shijiaxin3@huawei.com

4.5参考数据集

预训练数据：预训练数据网址：<https://huggingface.co/datasets/c4/tree/main>

使用 git clone <https://huggingface.co/datasets/c4/tree/main>进行下载，需要安装 git lfs。

下载完成后，选取其中的 c4/en/c4-train.00000-of-01024.json.gz、c4/en/c4-train.00001-of-01024.json.gz、c4/en/c4-train.00002-of-01024.json.gz、……、c4/en/c4-train.00013-of-01024.json.gz，共计14个文件，作为训练集。每个文件是一个压缩文件，解压后提取文本字段，大约735M。

下游任务数据：SST-2：<https://huggingface.co/datasets/SetFit/sst2/tree/main>

MNLI：<https://huggingface.co/datasets/SetFit/mnli/tree/main>

SQuADv1.1：<https://rajpurkar.github.io/SQuAD-explorer/dataset/train-v1.1.json>