CCKS2023 开放环境下的知识图谱构建与补全评测

# 任务简介

知识图谱是一种结构化的数据表示形式，它以图形的方式展示实体及其之间的关系。 而大型语言模型，如GPT系列模型，通过在大量文本数据上进行预训练，能够捕捉到丰富的语义和句法信息。 知识图谱可以为大型语言模型提供结构化的知识，有助于改善模型的推理能力和生成质量。通过将知识图谱中的信息融入大型语言模型中，可以使模型在处理复杂任务时更具智能和准确性。今年，面向更有挑战的开放环境下知识图谱构建和补全场景，围绕基于大模型的知识抽取与推理关键技术，浙江大学联合OpenKG开放知识图谱社区在CCKS2023组织评测并包括两个子任务：**指令驱动的自适应知识图谱构建、归纳式知识图谱关系推理**。两个评测任务具体内容在下面章节介绍。本次评测将依托天池平台（<https://tianchi.aliyun.com/>）展开，并鼓励评测选手使用大模型（如ChatGPT等），我们将持续维护和新增更多的评测任务，请有意向的参赛队伍请关注平台上的竞赛列表。

# 赛程安排

 本次大赛分为报名组队、初赛、复赛和决赛三个阶段，具体安排和要求如下：

 报名组队——————4月14日—7月12日

 初赛阶段——————4月19日—7月12日

 复赛阶段——————7月14日—7月21日

 评测论文——————8月11日

 评测颁奖——————8月24日-27日

 报名组队与实名认证（2023年4月14日—7月12日）

 报名方式：4月14日阿里天池平台(<https://tianchi.aliyun.com/>)将开放本次比赛的组队报名、登录比赛官网，完成个人信息注册，即可报名参赛；选手可以单人参赛，也可以组队参赛。组队参赛的每个团队2-3人，每位选手只能加入一支队伍；

 选手需确保报名信息准确有效，组委会有权取消不符合条件队伍的参赛资格及奖励；

 选手报名、组队变更等操作截止时间为7月12日23：59：59；各队伍（包括队长及全体队伍成员）需要在7月12日23：59：59前完成实名认证（认证入口：天池官网-右上角个人中心-认证-支付宝实名认证），未完成认证的参赛团队将无法进行后续的比赛；

 大赛官方钉钉群请搜索Group Number：27525018627，或扫描以下二维码加入，最新通知将会第一时间在群内同步：



钉钉群二维码

 初赛阶段（2023年4月19日-2023年7月12日，UTC+8）

 初赛的几个关键时间点： 4月19号天池平台将开放竞赛数据集和系统测评。

选手报名成功后，参赛队伍通过天池平台下载数据，本地调试算法，在线提交结果。

 初赛提供训练数据集，供参赛选手训练算法模型；同时提供测试数据集，供参赛选手提交评测结果,参与排名。初赛阶段提交格式在具体任务章节有详细介绍。

 初赛时间为2023年4月19日-2023年7月12日，系统每天提供2次评测机会，系统进行实时评测并返回成绩，排行榜每小时进行更新，按照评测指标从高到低排序。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示。

 初赛淘汰：2021年7月12日23：59：59，初赛阶段未产出成绩的队伍将被取消复赛参赛资格。

 初赛结束，初赛排名前100名的参赛队伍将进入复赛，复赛名单将在7月14日21：59：59之前公布。

 复赛阶段（2023年7月14日—2023年7月21日，UTC+8）

复赛阶段会提供复赛测试数据集下载，供参赛选手提交评测结果，参与排名。为防止作弊，复赛测试数据集可能会掺入无效评测数据。复赛阶段提交规范和初赛阶段保持一致。

 复赛时间为2023年7月14日-2023年7月21日。本阶段，系统每天提供5次实时评测，每小时更新排行榜，按照评测指标从高到低排序。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示。复赛提交截止时间7月21日17：59：59。
 复赛结束后，大赛组织方将通知top3优胜队伍提交代码（提交要求见具体的子任务章节），赛题组织方将对优胜参赛队伍代码进行审核，要求模型能复现出最优提交成绩。对于未提交、复现未成功或审核不通过的队伍，将取消决赛资格和比赛奖励，并通知递补选手。最终队伍排名按照复赛成绩核算，榜单将在官方完成代码审核后公布。

本次赛事鼓励获奖选手提交评测论文交流技术方案，其中录用的英文论文会被Springer收录论文集并EI检索。

# 规则

1. 所有参赛选手都必须在天池平台管理系统中注册，本次比赛的参赛对象仅限全日制在校大学生（本科、硕士、博士、博后均可）和企业员工；

2. 参赛选手需确保注册时提交信息准确有效，所有的比赛资格及奖金支付均以提交信息为准；

3. 参赛选手在管理系统中组队，参赛队伍成员数量不得超过3个，报名截止日期之后不允许更改队员名单；

4. 每支队伍需指定一名队长，队伍名称不超过15个字符，队伍名的设定不得违反中国法律法规或公序良俗词汇，否则组织者有可能会解散队伍；

5. 允许使用开源代码或工具，但不允许使用任何未公开发布或需要授权的代码或工具；

6. 参赛队伍可在参赛期间随时上传验证集的预测结果，初赛阶段每天2次机会，复赛阶段每天5次机会。管理系统会定时更新各队伍的最新排名情况。

7. 每名选手只能参加一支队伍，一旦发现某选手以注册多个账号的方式参加多支队伍，将取消相关队伍的参赛资格；

8. 评测任务均可使用外部数据；评测任务均可使用大模型，评测任务严禁人工标注数据，一经查实取消参赛资格。

9.浙江大学学生可以参赛和参与评奖。

10.评测任务规则解释权为评测组织方所有。

# 奖励设置

本任务设置一、二、三等奖和技术创新奖，均获得中国中文信息学会颁发的获奖证书，并有机会录用EI检索的评测论文。

# 任务组织者和联系人

**任务组织者：**

张宁豫，浙江大学

 张文，浙江大学

桂鸿浩，浙江大学

陈名杨，浙江大学

黄志伟，浙江大学

**学术指导委员会：**

陈华钧，浙江大学

熊飞宇，阿里巴巴

郭威，阿里巴巴

黄非，阿里巴巴

王昊奋，同济大学

漆桂林，东南大学

**联系方式：**

aliopenkg@service.alibaba.com

# CCKS2023 面向数字商务的知识处理与应用评测任务一：指令驱动的自适应知识图谱构建

### 任务描述

 指令驱动的自适应知识图谱构建是一种以用户指令为基础，实时更新和优化知识图谱结构的方法。这种方法的核心目标是根据用户需求调整知识图谱的表示，从而实现更准确、更高效的信息检索和推理，以满足开放环境下高效知识图谱构建的需求。在指令驱动的自适应知识图谱构建过程中，系统（如大模型）首先接收用户指令，识别其意图和需求。然后，根据指令内容，系统会基于输入的文本抽取相应的实体关系三元组以构建知识图谱。如果有知识缺失的情况，系统会根据用户指令中的补全需求实现对知识图谱的补全。这种方法的优点在于它可以实现知识图谱的定制计划构建。通过自适应调整，知识图谱能够更好地满足不同用户在不同场景下的需求。此外，指令驱动的自适应知识图谱构建方法还可以提高知识图谱的可扩展性，使其能够在面对新的数据和任务时进行有效适应。

### 任务目标

根据用户输入的指令抽取相应类型的实体和关系，构建知识图谱。其中10%指令包含了知识图谱补全，即任务需要模型在抽取三元组的同时对缺失三元组进行补全。

以下是一个**知识图谱构建任务例子**，输入一段文本text和instruction（包括想要抽取的关系类型），以“(ent1,rel,ent2)”的形式输出text中包含的所有关系三元组output\_text。

instruction = "请使用自然语言抽取三元组, 给定一个句子, 请从句子中抽取出可能的三元组, 需要抽取的关系类型为['父母', '配偶'], 请以[(头实体,关系,尾实体),...]的形式回答。"

text = "2008年6月23日，刘德华与朱丽倩在美国拉斯维加斯注册结婚。2012年5月9日，朱丽倩在香港养和医院产下女儿刘向蕙。刘向蕙出生后，少数看过刘向蕙面貌的圈内人士称，女儿的耳朵像刘德华，长得眉清目秀。"

output\_text = "[(刘德华,配偶,朱丽倩),(刘向蕙,父母,刘德华),(刘向蕙,父母,朱丽倩)]"

知识图谱补齐的含义是，在输入miss\_text（text中缺失了一段文字，text仍然通顺）和instruction的情况下，模型仍然能够补齐缺失的三元组，输出output\_text。下面是一个例子：

instruction = "请使用自然语言抽取三元组, 给定一个句子, 请从句子中抽取出可能的三元组, 需要抽取的关系类型为['父母', '配偶'], 请以[(头实体,关系,尾实体),...]的形式回答。"

miss\_text = "2012年5月9日，朱丽倩在香港养和医院产下女儿刘向蕙。刘向蕙出生后，少数看过刘向蕙面貌的圈内人士称，女儿的耳朵像刘德华，长得眉清目秀。"。

output\_text = "[(刘德华,配偶,朱丽倩),(刘向蕙,父母,刘德华),(刘向蕙,父母,朱丽倩)]"

虽然miss\_text中不包含“2008年6月23日，刘德华与朱丽倩在美国拉斯维加斯注册结婚。”这段文字，但是模型能够补齐缺失的三元组，即仍然需要输出(刘德华,配偶,朱丽倩),(刘向蕙,父母,刘德华)。

### 初赛训练集&测试集

**训练集**

在初赛训练集和验证集发布阶段，我们会发布5000条左右的知识图谱数据作为训练集。这5000条数据是使用维基数据对中文维基百科进行远程监督标注的弱监督数据，所以可能会存在一些错误。训练集包括一个train.json文件，可以在天池平台下载获取。

* train.json：每行是一个json串，具体字段内容如表所示

|  |  |
| --- | --- |
| 字段 | 说明 |
| id | 样本唯一标识符 |
| cate | 文本主题 |
| text | 模型输入文本（需要抽取其中涉及的所有关系三元组） |
| instruction | 模型进行抽取任务的指令 |
| output\_text | 模型期望输出，以(ent1,relation,ent2)形式组成的输出文本 |
| kg | text中涉及的知识图谱（由(头实体,关系,尾实体)三元组的格式组成） |

文本主题为以下12种类型之一：

1.人物（人、虚拟的人、名称等）

2.地理地区（城市、国家、河流、山脉、大洲、湖泊等）

3.事件（战争、活动、赛事等）

4.组织（企业、政府机构、机构、家族、队伍等）

5.生物（动物、植物、微生物、种等）

6.人造物件（材料、食物、设备等）

7.化合物

8.医学（健康问题、生理状况等）

9.运输（物流、航空、铁路系统、船运等）

10.建筑（建筑物、设施等）

11.天文对象（天体等）

12.作品（电影、数据、音乐等）

下面是一个json串示例：

**{**

 **"id": 0,**

**"cate: "运输",**

 **"instruction": "使**请使用自然语言抽取三元组, 给定一个句子, 请从句子中抽取出可能的三元组, 需要抽取的关系类型为[**'创建时间', '所在行政领土', '车站等级', '线路', '邮政编码']**, 请以[(头实体,关系,尾实体),...]的形式回答。**",**

 **"text": "沙河集站是一个京沪线上的铁路车站，位于安徽省滁州市南谯区沙河镇龙亭社区，建于1909年，目前为四等站，邮政编码为239060。目前客运：办理旅客乘降；行李、包裹托运；货运：办理整车货物发到；危险货物仅办理整车农药、化肥发到。",**

 **"output\_text": "[(沙河集站,线路,京沪线),(沙河集站,所在行政领土,龙亭社区),(龙亭社区,所在行政领土,沙河镇),(沙河镇,所在行政领土,南谯区),(南谯区,所在行政领土,滁州市),(滁州市,所在行政领土,安徽省),(沙河集站,车站等级,四等站),(沙河集站,创建时间,1909年),(沙河集站,邮政编码,239060),(京沪线,所在行政领土,龙亭社区)]",**

 **"kg": {**

**[**

**"沙河集站",**

**"线路",**

**"京沪线"**

**],**

**[**

**"沙河集站",**

**"所在行政领土",**

 **"龙亭社区"**

**],**

**[**

**"龙亭社区",**

**"所在行政领土",**

**"沙河镇"**

**],**

**[**

**"沙河镇",**

**"所在行政领土",**

**"南谯区"**

**],**

**[**

**"南谯区",**

**"所在行政领土",**

**"滁州市"**

**],**

**[**

**"滁州市",**

**"所在行政领土",**

**"安徽省"**

 **],**

**[**

 **"沙河集站",**

**"车站等级",**

**"四等站"**

 **],**

 **[**

 **"沙河集站",**

**"创建时间",**

**"1909年"**

 **],**

**[**

**"沙河集站",**

**"邮政编码",**

**"239060"**

 **],**

 **[**

 **"京沪线",**

**"所在行政领土",**

**"龙亭社区"**

 **]**

**}**

**}**

**初测试集赛**

在初赛阶段，我们会分两阶段（每阶段发布1000条）共发布2000条左右（含200条需补全） 的测试集。初赛测试集同样由一个json文件构成，每行是一个json串，格式和train.json一致，仅缺失kg、output\_text字段。

{

 **"id": 1,**

**"cate": "人物",**

 "instruction": "**使**请使用自然语言抽取三元组, 给定一个句子, 请从句子中抽取出可能的三元组, 需要抽取的关系类型为['职业', '父母', '国籍'], 请以[(头实体,关系,尾实体),...]的形式回答。",

 "text": "雷夫大叔冰岛人，原出生地不详，真实身份为传说中的挪威海盗红魔艾瑞克之子。",

}

### 复赛测试集

{

 **"id": 2,**

**"cate": "作品"**

 "instruction": "**使**请使用自然语言抽取三元组, 给定一个句子, 请从句子中抽取出可能的三元组, 需要抽取的关系类型为['杀青时间', '取景地点', '导演', '上映时间', '改编自', '产地', '开机时间', '主演'], 请以[(头实体,关系,尾实体),...]的形式回答。",

 "text": "欢乐元帅又名（春光灿烂之欢乐元帅）是2011年中国开拍的电视剧，翻拍自2000年电视剧《春光灿烂猪八戒》。以古装、神话、作题材，由福纳影视公司投资，著名作家范小天导演执导，古巨基、蔡卓妍、钟欣潼、郭晋安、吴卓羲、恬妞等人主演。此剧取景于江苏昆山福纳影视基地，于2011年4月28日开机拍摄。全剧共分四个单元，该剧共有50多位中港台明星组成的强大演员阵容；预计7月杀青。2012年1月全国电视台上映。",

}

在复赛阶段我们会发布1000条测试集（含100需补全） ，选手可以下载测试数据集。测试数据集由一个test.json文件构成，每行是一个json串，格式和valid.json一致。复赛结果提交方式和初赛保持一致，具体参见任务提交说明章节。

### 外部数据

 可以使用第三方数据，但是禁止人工标注数据。

### 评测指标

 本次任务采用F1值（F1-measure, F1）和ROUGE-N分数的加权和来评估模型进行知识图谱构建和补全的效果。精确率为在被所有预测为正的样本中实际为正样本的概率，召回率为在实际为正的样本中被预测为正样本的概率，F1值计算公式如下：

$$F1值=\frac{2\*精确率\*召回率}{精确率+召回率}$$

ROUGE-N将模型生成的output\_text和glod output\_text分别转换成一系列单独的单词，通过计算N-gram 的重叠程度来衡量它们之间的相似性，这里的N我们取2。所以最终score为：

score = 0.5 \* F1 + 0.5 \* ROUGE-N

注意，F1值计算的是预测的kg和gold kg之间的F1值，当预测的[pred\_head, pred\_rel, pred\_tail]与gold [head, rel, tail]完全相同时才算预测正确。

### 基线模型

本任务可以基于预训练语言模型如mT5 或大模型GPT-3, ChatGPT，GPT-4等。

方案不限于小项目微调，大模型指令学习、In-Context Learning， 大模型数据生成和增强等。

我们提供基线代码位于：

* mT5：<https://github.com/zjunlp/DeepKE/tree/main/example/triple/mt5>
* ChatGPT：<https://github.com/zjunlp/DeepKE/tree/main/example/llm>
* ModelScope：<https://modelscope.cn/models/damo/nlp_gpt3_text-generation_chinese-large/summary>

### 任务提交说明

比赛期间，参赛队伍通过天池平台下载数据，本地调试算法，在线提交结果，结果文件命名为"参赛队名称\_result.json"，以utf-8编码格式保存。文件每行是一个json串，包含"id"，"cate"，"text"，"output\_text"、"kg"字段，"kg"中每个三元组都按头实体、关系、尾实体的顺序排列。

例如：

**{**

**"id":3,**

**"cate": "运输",**

**"text": "沙河集站是一个京沪线上的铁路车站，位于安徽省滁州市南谯区沙河镇龙亭社区，建于1909年，目前为四等站，邮政编码为239060。目前客运：办理旅客乘降；行李、包裹托运；货运：办理整车货物发到；危险货物仅办理整车农药、化肥发到。",**

 **"output\_text": "[(沙河集站,线路,京沪线),(沙河集站,所在行政领土,龙亭社区),(龙亭社区,所在行政领土,沙河镇),(沙河镇,所在行政领土,南谯区),(南谯区,所在行政领土,滁州市),(滁州市,所在行政领土,安徽省),(沙河集站,车站等级,四等站),(沙河集站,创建时间,1909年),(沙河集站,邮政编码,239060),(京沪线,所在行政领土,龙亭社区)]",**

 **"kg": {**

 **[**

 **"沙河集站",**

 **"线路",**

 **"京沪线"**

 **],**

 **[**

 **"沙河集站",**

 **"所在行政领土",**

 **"龙亭社区"**

 **],**

 **[**

 **"龙亭社区",**

 **"所在行政领土",**

 **"沙河镇"**

 **],**

 **[**

 **"沙河镇",**

 **"所在行政领土",**

 **"南谯区"**

 **],**

 **[**

 **"南谯区",**

 **"所在行政领土",**

 **"滁州市"**

 **],**

 **[**

 **"滁州市",**

 **"所在行政领土",**

 **"安徽省"**

 **],**

 **[**

 **"沙河集站",**

 **"车站等级",**

 **"四等站"**

 **],**

 **[**

 **"沙河集站",**

 **"创建时间",**

 **"1909年"**

 **],**

 **[**

 **"沙河集站",**

 **"邮政编码",**

 **"239060"**

 **],**

 **[**

 **"京沪线",**

 **"所在行政领土",**

 **"龙亭社区"**

 **]**

 **]**

 **}**

**}**

模型不允许在测试集上进行训练，只能在训练集上训练。最终我们会对代码进行审核。

### 最终提交文件要求

 复赛结束后，组委将会联系获奖参赛队伍提交代码和模型：1）入口运行文件；2）代码及说明；3）方法描述文档。以上三个文件需在复赛结束后3天内发送至邮箱：aliopenkg@service.alibaba.com。邮件的标题为："天池-CCKS203任务1-最终提交文件-参赛队名称"。代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），用"code.指令KGC.参赛队伍名称"命名，要求提交所有的程序代码及相关的配置说明，确保程序能够正确运行，且所得结果与复赛结果相符。方法描述文档用"指令KGC.参赛队伍名称.pdf "命名，包含算法描述及参数设置，需用pdf格式存储，页数不超过10页。

# CCKS2023 开放环境下知识图谱构建与补全评测任务二：归纳式知识图谱关系预测

### 任务描述

知识图谱表示学习用于将符号表示的知识图谱映射到向量空间中，并且保留原本知识图谱中的语义信息。知识图谱表示学习当前已经被应用在大量的知识图谱相关的任务中，如链接预测、实体分类以及辅助各种知识图谱相关的下游任务。在开放环境下，知识图谱本身是动态变化的，并且新的知识图谱也在不断的构建当中，所以往往会有新的实体出现。但新实体的出现会让大多的传统知识图谱表示方法无法解决，因为他们大多假设测试阶段的实体在训练阶段见过，即直推式（transductive）设定，而很难直接应用到测试集中出现新实体的归纳式（inductive）场景。

因此，本子任务评估模型在测试阶段对新实体之间的关系进行预测。在训练阶段，会给定一个知识图谱用于训练（Training KG）；在测试阶段有两部分的三元组，且这些三元组中的实体在训练阶段中都没有出现过，分别称为支持集（Support Set）和查询集（Query Set）。支持集用于提供新出现实体的结构关系，最终模型在查询集上进行测试。

### 任务目标

在查询集中，给定一个头实体关系对（h, r），以及相应的尾实体的候选集合，模型输出候选集合的排序。候选尾实体越有可能是该（h, r）所连接的尾实体，排序越靠前。

|  |  |
| --- | --- |
| **输入** | **输出** |
| 头实体、关系对（h，r）；尾实体候选集合 | 尾实体候选集合排序 |

首先模型可以在训练数据集上进行训练，训练三元组由编号（index）表示，例如：

0 0 1

2 1 3

4 2 5

6 3 7

8 4 9

10 5 11

其中每一行表示一个三元组，每一行三个数字分别表示训练集中的头实体、关系、尾实体的编号。

我们也提供了验证集合，用于对模型训练的验证

### 数据集介绍

### 训练数据集合

训练集：train\_triple.txt，每行为一个三元组（head, relation ,tail），其中用数字表示实体和关系的编号。实体和关系编号分别从0开始。

### 测试数据集合

支持集：test\_support.txt，每行为一个三元组（head, relation ,tail），其中用数字表示实体和关系的编号。这里的关系编号和训练集合是一致且对应的，但实体编号并不对应。测试集中的实体都是训练集中没见过的新实体，新实体的编号也从0开始。

查询集：test\_query.txt，每行前两个数字为头实体和关系，后面的数字为候选尾实体集合。这里的实体编号和支持集中的一致。

### 评测指标

 本次任务采用MRR评估对尾实体的预测，MRR值越高，说明对正确尾实体的排序越高，预测越准确。MRR的定义如下：

$$MRR=\frac{1}{N}\sum\_{}^{}\frac{1}{rank\_{N}}$$

其中N是测试集中查询集合的大小，$rank\_{N}$ 为第N个测试头尾实体对所预测的候选尾实体排序中，正确尾实体的排序位置。

### 基线模型

Baseline模型采用基于子图采样的方法GraIL，该方法将通过编码头实体和候选尾实体之间的子图结构，从而预测头实体和某个候选尾实体之间是否存在具体的关系。

### 任务提交说明

 比赛期间，参赛队伍通过天池平台下载数据，本地调试算法，在线提交结果，结果文件命名为"参赛队名称\_result.txt"，以utf-8编码格式保存。每行是测试集的查询集中每个头尾实体对应的候选尾实体的排序结果，以逗号进行分隔。

### 模型不允许在测试集的支持集上进行训练，只能在训练集上训练。最终我们会对代码进行审核。

### 最终提交文件要求

 复赛结束后，组委会将联系获奖参赛队伍提交代码和模型：1）入口运行文件；2）代码及说明；3）方法描述文档。以上三个文件需在复赛结束后3天内发送至邮箱：aliopenkg@service.alibaba.com。邮件的标题为："天池-CCKS2023任务2-最终提交文件-参赛队名称"。代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），用"code.常识显著性.参赛队伍名称"命名，要求提交所有的程序代码及相关的配置说明，确保程序能够正确运行，且所得结果与复赛结果相符。方法描述文档用"常识显著性.参赛队伍名称.pdf "命名，包含算法描述及参数设置，需用pdf格式存储，页数不超过10页。相关代码建议上传到代码托管平台。