**CCKS 2022 技术评测任务书**

# 跨语言知识图谱问答

# 任务描述

**随着近年来越来越多的非英语使用者参与到知识图谱的建立和维护，在线知识的分布已经由富资源语言(英语)一家独大转变为多语言资源相互补充。然而，当前多语言知识图谱问答模型主要关注于自然语言问题解析，而忽略了对跨语言知识的联合应用。**

**以百科知识问答为例，对于新出现的知识，如**“**COVID-19 (2019新冠病毒病)**”**，在Wikipedia中文词条中，关于“并发症”属性的记录包含**“**全身器官衰竭**”**，而这项记录未被收录于英文词条关于**“**Complication(并发症)**”**的条目中。另一方面，在英文词条中包含了**“**Duration(症状持续时间)** ”**的属性，而中文词条却缺乏对这项属性的收集。因此，诸如：**“**能够引起全身器官衰竭这一并发症的冠状病毒，它使人体产生的症状一般会持续多长时间**？”**，**“**症状一般持续五天以上，且可能转为慢性的冠状病毒，是否可能引发全身器官衰竭这类严重并发症**？”**，对于这类需要联合跨语言知识的问题，目前的问答系统难以回答。**

此次评测任务的目标是利用跨语言知识图谱回答不同语言提出的问题**，**我们期望参赛选手的问答系统既能够解决自然语言问题与知识图谱间的词法鸿沟问题, 又能在跨语言知识图谱中准确找到问题相关的知识组合**。**

* **输入**

输入文件包含若干行自然语言问句。

* **输出**

输出文件每一行对应一个问题的答案路径，列表内元素以\t分隔，路径元素由分隔符#隔开。

* **输入样例**

**q1:** 伊朗裔美国艺术家和教育家Taravat Talepasand的母校附属于哪个机构？

**q2:** **Who does the person whose name is used to name Poe (crater) influence?**

* **输出样例**

**a1:**

<http://dbpedia.org/resource/Taravat\_Talepasand>#

<http://dbpedia.org/property/almaMater>#

<http://dbpedia.org/resource/Rhode\_Island\_School\_of\_Design>#

<http://zh.dbpedia.org/resource/罗德岛设计学院>#

<http://zh.dbpedia.org/property/state>#

<http://zh.dbpedia.org/resource/罗德岛州>

**a2:**

<http://dbpedia.org/resource/Poe\_(crater)>#

<http://dbpedia.org/property/eponym>#

<http://dbpedia.org/resource/Edgar\_Allan\_Poe>#

<http://zh.dbpedia.org/resource/爱伦·坡>#

<http://zh.dbpedia.org/property/influenced>#

<http://zh.dbpedia.org/resource/江户川乱步>

# 数据描述

* **数据来源：**

本次评测的问答数据由模板构建和自动复述生成方式建立，其中模板由来自东南大学，江西师范大学和九江学院的汉语，英语和法语学科的本、硕、博同学参与标注。自动复述生成方法基于东南大学KSE实验室QA组提出的KTG问题生成模型搭建 (发布于COLING 20)，通过引入Wikidata外部知识为生成问题构建复述。问题集经过对应语言专业的同学人工验证及后处理，能够保证数据的准确性和多样性。问答数据集由中，英，法三语平行的路径问题(path question)构成，包含2-hop和3-hop两种问题类型(对应答案路径包含2个或3个三元组)。

* **训练集&验证集：**

在训练数据发布阶段，我们会对发布15000条标注好的数据(包括问题/答案路径，每种语言5000条)，作为训练集。同时发布1500条左右不含标注的问题作为验证集（每种语言500条），选手可以将自己生成的答案路径提交，比赛系统会对答案进行评测，给出得分并进行排序。在测试数据发布阶段，我们会发布验证集的标注结果(包括问题/答案路径)，同时发布1500条左右不含标注结果的问题，作为测试集。

* **知识库相关文件说明**

本评测任务使用给定知识图谱，该图谱采样自DBpedia (2016-10)，分为“中-英”，“英-法”，“中-法”三组跨语言知识图谱，每组对应有独立的问题集，并提供Inter language link (ILLs) 资源作为跨语言知识图谱间的实体对齐标注。数据说明（如有变动后续会进行更新，以实际下载文件为准）：

1. Triples\_zh.txt包含中文知识图谱的主要三元组。
2. Triples\_en.txt包含英语知识图谱的主要三元组。
3. Triples\_fr.txt包含法语知识图谱的主要三元组。
4. ILLs\_zh\_en.txt包含“中-英”跨语言图谱的已知实体对齐。
5. ILLs\_en\_fr.txt包含“英-法”跨语言图谱的已知实体对齐。
6. ILLs\_zh\_fr.txt包含“中-法”跨语言图谱的已知实体对齐。

# 评价指标

本任务的评价指标为问答的精准率(Accuracy), 最终排名以Accuracy为基准。设为问题集合,为选手对第个问题给出的答案路径包含的有序节点集合, 为第个问题的标准答案路径包含的有序节点集合,相关计算方式如下:

**精准率(Accuracy)**: 当答案路径与标注路径的节点完全一致时表示回答正确

**注：**经过筛选，当前版本的任务数据不会出现一个问题对应多条答案路径的情况。

# 任务提交

**提交方式：**

本次任务将采取刷榜的方式，分为两个阶段进行：

第一阶段：验证集发布后，允许参赛队伍多次向平台提交结果。参赛者将所有预测结果写入一个结果文件中上传给系统（每日最多可上传3次），如果新提交结果好于之前提交结果，排行榜中的成绩将自动进行更新覆盖。

第二阶段：测试集发布后，验证集提交通道将会关闭，其余事项与第一阶段相同。

**最终提交文件要求：**

每一个参赛队需提交的材料如下。

* 1. 问答任务测试集结果文件，用result.txt命名（UTF-8格式）
  2. 相关代码及说明
  3. 方法描述文档（非评测论文，评测论文撰写要求见CCKS 2022官网）

以上三个文件需在任务提交截止日期前发送至邮箱tt\_yymm@seu.edu.cn（暂定）。邮件的标题为：“CCKS-CLKGQA-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.xxx命名，要求提交所有的程序代码及相关的配置说明，**程序应当可以运行且所得结果与result.txt相符**。如果方法使用了额外资源，要求说明并提供资源文件或地址。

**评测平台：**本次评测将依托biendata平台（https://biendata.com/）展开，请有意向的参赛队伍关注平台上的竞赛列表。

# 时间安排

* 任务征集截止：3月24日
* 任务准备时间：3月25日—4月6日
* 评测任务发布：4月6日
* 报名时间：4月6日—7月25日
* 训练及验证数据发布：4月25日
* 测试数据发布：7月25日
* 提交测试结果：7月31日
* 评测论文提交：8月12日
* CCKS会议日期(评测报告及颁奖)：8月25日—28日

# 规则

1. 报名方式：参赛选手需要提交“参赛队名，队长信息（姓名，邮箱，联系电话），参赛单位名称”等信息，发送至邮箱nanhu@seu.edu.cn，并在评测网站报名。报名截止到测试数据集发布，在测试数据集发布之后，未报名的选手/队伍不能再报名或提交。

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. 参赛队伍成员数量不得超过5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. 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；

6. 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；

7. 除主办方提供的数据集外，参赛选手允许使用预训练数据（如词向量、字向量等），但是不能用额外的标注数据；

# 任务组织者

漆桂林 （东南大学） gqi@seu.edu.cn

谭亦鸣 （东南大学） tt\_yymm@seu.edu.cn

胡楠 （东南大学） nanhu@seu.edu.cn

李嘉琦 （东南大学） 494296118@qq.com

# 奖项安排

第一名：冠军证书

第二名：亚军证书

第三名：季军证书

4-8名：优胜证书

本次任务将会选出一支队伍参与CCKS评测任务创新奖评选（奖金5000元）