### 开放领域的知识图谱问答评测

#### 评测背景

在互联网软硬件相关技术飞速发展的今天，人们每天接触与制造的数据量日益上升，理解与应用这些数据所需的成本也随之增加。因此，如何高效而准确地处理海量异质数据成为了一个亟待解决的问题。知识图谱以结构化的“知识”来存储与表示海量数据，作为承载底层海量知识并支持上层智能应用的重要载体，它在智能时代中扮演了极其重要的角色。

然而，由于知识图谱高度结构化的特点，我们常常需要构建结构化查询语句（SPARQL等）来查找相关知识，这为普通用户使用知识图谱造成了不便。因此，在知识图谱上进行自然语言问答（KBQA）近年来成为了前者的热门应用之一。在学界，semantic parsing、IR等创新性方法与框架百花齐放；在业界，智能音箱、语音助手、智能问诊等应用也极大地拓宽了知识图谱自然语言问答的应用场景，进一步加强了对高效、准确、易用、安全、可解释的KBQA系统的需求。

面对这一需求，我们提出了基于知识图谱的开放领域问答这一评测，期待参赛者们可以提出创新性的KBQA系统，同时处理“专而深”的特定领域和“广而浅”的通用领域知识图谱，对用户提出的复杂多样的自然语言问题给出准确答案。具体而言，我们的“开放”体现在两个方面：其一是我们使用的知识图谱数据融合了来自百度百科的通用领域知识和来自美团、恒生、妙健康等合作伙伴的专门领域知识。我们还对这些知识图谱进行润色，合并其中的公共实体，消除逻辑上互相矛盾的关系，使它们有机结合。其二，我们不仅为针对单个知识图谱编写了问答数据，而且编写了一部分需要从多个知识图谱中获得对应知识来解答的自然语言问题。

“开放领域”的特性是本次评测与其他基于知识图谱的问答评测的主要区别。据我们调查发现，此前的基于知识图谱的问答评测有很大一部分是针对特定领域的知识图谱设计的，而且用于评测的问题中简单问题占大多数，使得参赛者可以容易地将性能指标提高到95%以上。而我们提出的开放领域知识图谱问答融合了多个领域的知识图谱，同时在出题时有意提高了题目的复杂程度，从而要求参赛者的模型有更好的泛化能力以及处理复杂问题的能力，不仅能够应对涉及单一领域的问题，也能够从多个领域中抽取有用的信息，以解答跨领域的问题。

我们的团队有过举办类似评测的经验，对涉及的技术和流程较为熟悉。与我们此前举办的评测相比，我们沿用了评测的流程安排和性能指标，但是我们对知识图谱数据和问答数据都进行了更新。在之前的基础上我们使用了全新的验证集和测试集数据，并且对知识图谱数据的部分进行了实体对齐和歧义消除等优化。

需要特别强调的是，众所周知，最近横空出世的ChatGPT，以及后续跟进的大量GPT系列的工具，将自然语言对话与问答的效果提升到了新的高度。我们在内部调研时发现，ChatGPT在通用领域的效果较好，但在专用领域ChatGPT的能力还有待提高，很多时候它会输出看似头头是道实则为谬误的结果，它并不是一个100%可靠的知识库。我们认为这主要是由于在某些领域ChatGPT的训练语料并不完善导致的。而我们的评测恰恰着重考核模型同时应对通用领域和专业领域的自然语言问题的能力。除此之外，目前ChatGPT在实时性上还存在提升的空间，由于训练语料的限制，ChatGPT的回答无法马上同步最新的知识，需要在后台使用新的数据进行模型的迭代，或者将模型与搜索引擎等API结合，让其自动从外界获取新的知识。这两种方式的缺点，一是对知识图谱进行修改的代价无疑比模型本身进行迭代的代价更低，二是模型自发从搜索引擎等第三方API获取到的知识与模型本身训练使用的数据相比，其质量得不到保证。我们的评测更强调泛化能力，因此在知识图谱更新后，系统也能够输出正确的结果。无论如何，ChatGPT系列模型作为目前自然语言处理乃至人工智能研究中的掌上明珠，我们都不应该忽视，但也不能神化它。我们的团队在本次评测中也希望选手能从大规模语言模型的角度出发，提出自己的解决方案。

#### 任务描述

本任务属于开放领域的中文知识图谱自然语言问答任务，简称CKBQA （Chinese Knowledge Base Question Answering）。即输入一句中文问题，问答系统从给定知识库中选择若干实体或属性值作为该问题的答案。问题均为客观事实型，不包含主观因素。理解并回答问题的过程中可能需要进行实体识别、关系抽取、语义解析等子任务。这些任务的训练可以使用**额外的公开的**语料资源，但是最终的答案必须来自给定的知识库。

知识图谱问答在当前互联网信息爆炸、人工智能盛行的时代是十分有战略价值和研究意义的。一方面，传统搜索引擎是以网页资源为核心，依据关键词索引、文本匹配等方式进行检索并返回给用户相关网页链接，而用户很多时候需要的只是对一个具体问题的特定解答。另一方面，诸如智能音箱、智能问诊等新一代产品应用通常也需要依赖于特定的知识图谱响应用户的自然语言请求。例如智能问诊应用可以依据医药健康领域的知识库对患者的情况进行初步诊断。

本次知识图谱问答任务是在CCKS上举办的第五届。回顾以往几届，我们先后与恒生、妙健康、美团等业界代表合作，扩充了我们的知识库和问题集。今年，我们保留去年开放领域的问答数据以供参赛队伍对模型进行训练，同时我们还更新了验证集和测试集。我们期望参赛选手的问答系统既能处理各种百科类的浅层问题，也能处理具备一定领域知识。

本评测任务所使用的知识图谱以开放领域知识库PKUBASE为基础，先后整合了金融、医药和生活旅游领域的知识图谱。

#### 输入输出

* **输入**

输入文件包含若干行中文问句。

* **输出**

输出文件每一行对应一个问题的答案列表，列表内元素以制表符\t分隔。

* **输入样例**

q1: 故宫附近有哪些豪华酒店？

q2: 紫竹院公园地址在哪？

q3: 朝阳区有哪些不收门票的公园？

* **输出样例**

<北京王府井希尔顿酒店>\t<北京励骏酒店>\t<北京国际饭店>\t<北京东方君悦大酒店>

<北京市海淀区中关村南大街35号>

<奥林匹克森林公园>\t<将府公园>\t<黑桥公园>

* **说明**

为帮助参赛选手提高系统性能，训练文件会提供问题对应的SPARQL查询。

例句一对应的SPARQL查询语句如下：

 SELECT ?x WHERE

 {

<故宫> <附近>    ?x.

?x <类别> <酒店>.

?x <等级> <豪华>.

}

例句二对应的SPARQL查询语句如下：

 SELECT ?x WHERE

{

<紫竹院公园> <地址>    ?x.

}

例句三对应的SPARQL查询语句如下：

 SELECT ?x WHERE

{

?x <类型> <公园>.

?x <行政区> <朝阳区>.

?x <门票价格> <免费>.

}

SPARQL语言的语法规则可以参考<https://www.w3.org/TR/rdf-sparql-query/>。原则上不要求生成SPARQL查询，参赛选手可以选择自己的方式，只需要给出问题的最终答案（答案需要来自给定知识库）。**当问题所涉及的事实与知识库不一致时，以知识库中数据为准。**

输出答案中由<>括起的为实体，由""括起的为文本值。请严格按照问题顺序进行答案输出，若参赛系统对某一问题产生答案为空，请输出空行。

#### 评价指标

本任务的评价指标包括宏观准确率(Macro Precision)，宏观召回率(Macro Recall)，Averaged F1值。最终排名以Averaged F1值为基准。设$Q$为问题集合，$A\_{i}$为选手对第$i$个问题给出的答案集合，$G\_{i}$为第$i$个问题的标准答案集合，相关计算公式如下：

$Macro Precision=\frac{1}{|Q|}\sum\_{i=1}^{|Q|}P\_{i}$, $P\_{i}=\frac{|A\_{i}\bigcap\_{}^{}G\_{i}|}{|A\_{i}|}$

 $Macro Recall=\frac{1}{|Q|}\sum\_{i=1}^{|Q|}R\_{i}$, $R\_{i}=\frac{|A\_{i}\bigcap\_{}^{}G\_{i}|}{|G\_{i}|}$

$$Averaged F1=\frac{1}{|Q|}\sum\_{i=1}^{|Q|}\frac{2P\_{i}R\_{i}}{P\_{i}+R\_{i}}$$

#### 数据集

* **数据来源**

本次测评的问答数据来自于人工构建与标注。标注过程不依赖于特定模板，标注人员超过20人以保证问题各方面的多样性。问答数据集中的问题不局限于特定领域，且既包含简单问题（对应单元组查询），也包含复杂问题（对应多元组查询），它们的数量比例大致为1:1。

* **训练集&验证集**
* 在训练数据发布阶段，我们会发布9000条标注好的数据（包括问题/SPARQL/答案）作为训练集。同时发布900条左右不含标注结果的问题作为验证集，选手可以将自己生成的验证集答案提交，比赛系统会对答案进行评测，给出得分并进行排行。此时的排行仅作为参考。

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/SPARQL/答案），同时发布900条左右不含标注结果的问题，作为测试。测试集上的最终排名确定后

* **知识库相关文件说明**

本任务使用的知识库主要由开放领域知识库PKUBASE及多个领域数据集融合得到。参赛选手可以在beindata平台下载数据文件后使用相应的知识库管理系统（例如gStore系统：<http://gstore.cn/> ）进行存储和查询。同时，为方便参赛选手完成任务，我们也提供在线查询终端，选手可以通过浏览器或调用API进行SPARQL查询。详情访问<http://pkubase.gstore.cn/>。

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.txt包含知识库主要三元组。
2. types.txt包含各实体的类别三元组。
3. mention2ent.txt可以用来辅助选手进行实体链接。其格式为“短语\t候选实体\t候选实体的排名”，如下例短语“逆时针”排名第一的候选实体是逆时针\_（汉语名词）。此文件仅供参考，其覆盖范围及准确度并无保证，选手可以视情况使用。**如需对此映射文件进行增补，请在文档中写明增补的策略并提交相关的代码，原则上不允许人工根据测试问题添加mention2ent映射。**

逆时针 逆时针\_（汉语名词） 1

逆时针 逆时针\_（张靓颖演唱歌曲） 2

逆时针 逆时针\_（化妆品品牌） 3

…

#### 任务提交

本次评测将采取刷榜方式，各任务验证集发布后，允许参赛队伍多次向平台提交结果，文件命名参考具体任务说明，并以队伍名作为前缀。格式与任务描述中的示例输出相同（注意格式，如对于问答任务需要保留<>或""），排名实时更新。参赛队伍可在评测集发布之前随时上传验证集的计算结果，管理系统会及时更新各队伍的最新排名情况；

测试集发布后，允许参赛队伍多次提交测试集结果文件（每天提交不超过3次）。

**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linyinnian@pku.edu.cn。邮件的标题为：“CCKS-CKBQA-参赛队名称”，例如“CCKS-CKBQA-火箭队”。

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。

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。

#### 数据集

评测任务发布：3月31日

报名时间：4月1日 - 7月14日

训练及验证数据发布：4月14日

测试数据发布：7月14日

最终测试结果：7月21日

评测排名通知：7月28日

评测论文提交：8月11日

CCKS会议日期（评测报告及颁奖）：8月24日-27日

#### 评审规则

1. 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，报名方式稍后在评测网站发布。
2. 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。
3. 每支队伍需指定一名队长，队伍名称不超过15个字符，队伍成员不超过4人。
4. 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，将取消所有相关队伍的参赛资格。
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。允许使用外部数据，但该数据必须是公开的，并在提交最终结果时一并提交（如数据过大，需提供下载地址）。
6. **鼓励使用通用的、创新的算法、模型解决问题，不允许基于测试集或KB用规则Case By Case解决问题，若发现主办方有权取消参赛资格。**
7. **参赛选手最终需要提交可运行的代码和方法描述文档，若在排行榜上的结果无法复现，将取消参赛资格。**
8. 欢迎国内外在校生及社会在职人士参加。比赛组织方成员不可参赛。

#### 奖励规则

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。

# 组织者

邹磊 北京大学王选计算机研究所 zoulei@pku.edu.cn

林殷年 北京大学王选计算机研究所 linyinnian@pku.edu.cn

 张旻昊 北京大学王选计算机研究所 zhangminhao@pku.edu.cn

 张若禹 北京大学王选计算机研究所 ry\_zhang@pku.edu.cn