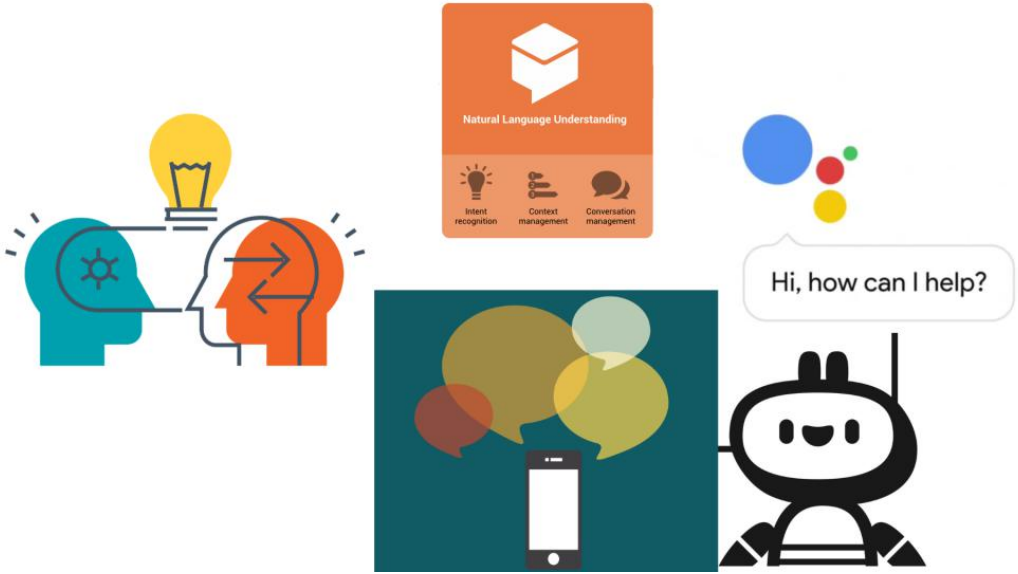
2020-人工智能基础课程大作业

中国科学院计算技术研究所

中科院智能信息处理重点实验室

作业内容

- 冬奥会领域问答机器人



提交内容与评分标准

- 以小组为单位完成Project，题目2选1，也可以都选（有额外加分）
  - 每组不超过4人，设组长一人
  - 组内分工不限，自行安排，但在最终设计报告中需要体现组内分工
- 评分标准
  - Project满分为100分，根据相应比例计入总成绩（比例为20%）
- 提交内容
  - 模型源代码
  - Project文档报告（中英文都可以，鼓励写英文）
- 报告最晚提交时间，最后一次课的前一周

问答大作业问题概述

- 概述：有关“***冬奥会比赛***”的问答系统，当输入一个有关冬奥会的问题，系统给出相应的答案。

- 例子：

问题: 哪一届冬奥会是亚洲举办的第一届奥运会?

答案:

(1) 第十一届日本札幌冬奥会 ✔
(2) 第十一届 ✔
(3) 11 ✔
(4) 日本札幌冬奥会 ✔

→ 返回任意一个正确答案即可

数据提供1—16k标注好的问答对(excel)

问题	答案	#1问句类型	#1领域类型	#1语义类型	#2问句类型	#2领域类型	#2语义类型
1924年在法国夏蒙尼冬奥会上决出的第一个项目是什么?	男子500米速度滑冰	Which	Competition	Factoid	NA	NA	NA
1924年夏慕尼冬奥会举办的时间是什么时候?	1924年1月25日至2月4日	When	Competition	Calculation	NA	NA	NA
1924年夏慕尼冬奥会的第一个比赛项目金牌获得者是谁?	查尔斯·朱特劳	Which	Competition	Factoid	Who	Competition	Factoid
1924年夏慕尼冬奥会的参赛运动员有多少人?	258人	How	Competition	Calculation	NA	NA	NA
1924年夏慕尼冬奥会的开幕式主持人是谁?	加斯东维达尔	Who	Competition	Factoid	NA	NA	NA

问题分层

a)单层问题

问句：*哪一届奥运会的金牌总数最多?*

分词：哪 一届 奥运会 的 金牌 总数 最 多 ？

#1问句类型：Which多选一

#1领域类型：Competition比赛

#1语义类型：Calculation计算

#2问句类型：NA

#2领域类型：NA

#2语义类型：NA

(b)叠加（双层）问题

问句：*中国奥运第一人值得尊敬吗?*

分词：中国 奥运 第一 人 值得 尊敬 吗 ？

#1问句类型：Who

#1领域类型：Competition

#1语义类型：Factoid

#2问句类型：Whether

#2领域类型：Competition

#2语义类型：Opinion

数据提供1—16k标注好的问答对(excel)

问句类型

类型	含义	举例
事实 (factoid)	询问确切答案	里约奥运会开幕式的时间?
描述性 (descriptive)	询问属性或因由等	里约的城市环境如何?
过程性 (procedural)	询问一种过程	护照如何办理?
计算 (calculation)	需要计算回答	羽毛球男单有几支队伍参赛?
推理 (inference)	需要分析、推断和预测	韩国队为何能出现?
观点性 (Opinion)	询问态度、情感和断言	你对罗伯特有何看法?

问题分类

类型	举例
When	短道速滑的比赛是哪一天举行?
Where	2012年奥林匹克运动会在哪举办?
Whether	孙杨是否参加100米自由泳决赛?
Who	奥运会开幕式的导演是谁?
Which	哪一届奥运会的金牌总数最多?
Why	为什么第24届奥运会我们的金牌数那么少?
How	怎么到达奥运会游泳馆?

数据提供2—知识图谱

- 知识图谱详见：https://xlore.org/?lang=cn
- Online使用（可离线使用提供的 WinterOlympics.zip）：

```
In [1]:
1 import requests
2 import json

In [2]:
1 req = requests.get('http://api.xlore.org/query?word=计算所')

In [3]:
1 data = json.loads(req.text)

In [4]:
1 data

Out[4]: {'Classes': [{'Hypernym': [],
   'Hyponymy': [],
   'Instances': [{'Label': '中国科学院计算技术研究所',
     'Uri': 'http://xlore.org/instance/zhi168560'},
    {'Label': '计算机体系结构国家重点实验室', 'Uri': 'http://xlore.org/instance/zhi54753'}],
   'Label': '中国科学院计算技术研究所',
   'SameAs': [{'Label': 'Chinese',
     'Uri': 'http://xlore.org/concept/zhc121700'}],
   'Uri': 'http://xlore.org/concept/zhc121700'}],
 'Instances': [{'Abstracts': "'中国科学院计算技术研究所'（简称“'中科院计算所'”）是[[zh3924]]下属的[[zh5]]技术研究机构，成立于1956
年，是中国第一个专门从事计算机领域研究的专门机构。所址位于[[zh128]][[zh11702]][[zh32059]]，并建有苏州、上海、烟台等多个分所（分部）。
现任所长为孙凝晖研究员。计算所有[[zh54753]]、中国科学院智能信息处理重点实验室、中国科学院网络数据科学与技术重点实验室等13个研究实体。计算
所自1960年起开展[[zh20900]]教育，是中国首批[[zh294]]、[[zh13695]][[zh92665]]授予点。目前在读研究生人数约为1000人。计算所曾陆续分离
出[[zh246803]]、[[计算中心]]、[[软件研究所]]和计算机网络信息中心等多个研究机构，以及[[联想]]、曙光、[[zh66367]]等高技术企业。",
   'Label': '中国科学院计算技术研究所',
   'Related Instances': [{'Label': '中国科学院软件研究所',
```

输入输出数据接口要求

- 输入、输出及要求：

1. 输入输出分别以json格式 {“answer”: “…”, “question”: “…”}

   例如：{“question”:“2022年冬奥会举办地是哪里”, “answer”:“北京”}。

2. 单次响应时间在500ms以内（除去初始加载模型的时间）。
3. 最终成绩评测：测试语料上返回正确答案的可接受率（提供的数据集上的随机划分，测试集上的性能）。（建议进行7/1/2，8/1/1等划分）

- 提示：

1. 系统模型以「检索式对话系统」或者「检索+生成式对话系统」为主
2. **可选择性使用**所提供的训练语料和知识图谱，也可自己**增加冬奥会对话数据集**（若使用了额外的数据集，需一并提交，考虑会给加分）。